CROP DECISION PLANNING UNDER YIELD AND PRICE UNCERTAINTIES

A Thesis
Presented to
The Academic Faculty

by

Nantachai Kantanantha

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2007

CROP DECISION PLANNING UNDER YIELD AND PRICE UNCERTAINTIES

Approved by:

Dr. Nicoleta Serban, Co-Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Paul M. Griffin, Co-Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kwok-Leung Tsui
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Gunter P. Sharp
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Steven Y. Liang
School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: June 14, 2007

To my mother ...

ACKNOWLEDGEMENTS

First of all I would like to thank my family. To my parents, I thank their endless love, support, and encouragement. Even though my mother is not here anymore, I always have her in my heart and I am sure that she would be proud of me. To the rest of my family - Num, Mee, Koi, P'Nuch, P'Poom, Win, and Mint - thank you for always being there.

I am deeply indebted to my advisors, Dr. Nicoleta Serban and Dr. Paul M. Griffin for their motivation, support, and guidance during my study at Georgia Tech. Without their help, I would not have come this far. Their knowledge, experience, and insights have been very influential in my research.

I would also like to thank Dr. Kwok-Leung Tsui, Dr. Gunter P. Sharp, and Dr. Steven Y. Liang. I appreciate their time and effort in serving on my dissertation committee.

I would like to give a special thanks to my girlfriend, Jing, for her love, encouragement, and companionship. Thanks for always being here for me. I thank my cousin, Jan, for her support and encouragement. I also thank my relatives, uncle Kai, aunt Paradee, Jeab, Kitti, and Kajorn for their warm hospitality when I visited them in West Virginia.

I express my special gratitude to my best friend here, Dr. Tiravat Assavapokee (To) for his support, friendship, and suggestion. Thanks also go to my friends at Georgia Tech. Thank you Kaulad, Kong, Auey, Kat, Orn, Winny, Josh, Nan, Nick, Tap, Golf, Pang, Chompoo, Noon, P'Kaew, P'Gnn, P'Tan, P'Oat, Nuch, Mam, Lynk, Brad, Jennifer, Leanne, and David. I am also thankful to my dinner gang - Oh, Chat, Gib, Nu, Chi, Tam, Ter, Term, Champ, and L for having dinner with me every friday.

I would also like to thank my friends in Thailand - Nares, Tongkamol, Pakdee, Boworn, Hor, Ple, Pop, Wi, P'Ping, Orm, Rong, Kobby, Nok, Som, Pook, X, Gift, Tui, and Thip for their friendship and understanding.

Lastly, I would like to thank other people who deserve the gratitude but I have not mentioned here. To them, please accept my apology and thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This research focuses on developing a crop decision planning model to help farmers make decisions for an upcoming crop year. The decisions consist of which crops to plant, the amount of land to allocate to each crop, when to grow, when to harvest, and when to sell. *The objective is to maximize the overall profit subject to available resources under yield and price uncertainties.*

To help achieve this objective, we develop yield and price forecasting models to estimate the probable outcomes of these uncertain factors. The output from both forecasting models are incorporated into the crop decision planning model which enables the farmers to investigate and analyze the possible scenarios and eventually determine the appropriate decisions for each situation.

This dissertation has three major components, yield forecasting, price forecasting, and crop decision planning. For yield forecasting, we propose a crop-weather regression model under a semiparametric framework. We use temperature and rainfall information during the cropping season and a GDP macroeconomic indicator as predictors in the model. We apply a functional principal components analysis technique to reduce the dimensionality of the model and to extract meaningful information from the predictors. We compare the prediction results from our model with a series of other yield forecasting models. For price forecasting, we develop a futures-based model which predicts a cash price from futures price and commodity basis. We focus on forecasting the commodity basis rather than the cash price because of the availability of futures price information and the low uncertainty of the commodity basis. We adopt a model-based approach to estimate the density function of the commodity

basis distribution, which is further used to estimate the confidence interval of the commodity basis and the cash price. Finally, for crop decision planning, we propose a stochastic linear programming model, which provides the optimal policy. We also develop three heuristic models that generate a feasible solution at a low computational cost. We investigate the robustness of the proposed models to the uncertainties and prior probabilities. A numerical study of the developed approaches is performed for a case of a representative farmer who grows corn and soybean in Illinois.

CHAPTER 1

INTRODUCTION

Decision planning plays an important role in agriculture as it does in other industries. It is a key factor that determines the success or failure of business. In this dissertation, we will focus on decisions made during the crop planning periods. The decisions that farmers have to make include

- Which crops to grow;
- What amount of the land to allocate to each crop;
- When to grow, when to harvest, when to sell.

However, it is difficult to make the right decisions. This is because the farmers have to take into account uncertain factors such as weather, demand, and supply as well as resource limitations. These factors result in uncertainties in yield and price, which significantly affect the return to producers. All of these problems, as shown in Figure 1, challenge the involved parties in determining a solution that will help farmers reach their goals.

A decision planning model is developed to establish a solution for this problem by taking into account the resource limitations as its major constraints. In order to make decisions under uncertainty, forecasting of uncertain factors is a crucial step since it can estimate the probable outcomes of the final output. Forecasting processes are established to estimate the values and confidence intervals of the uncertain or stochastic variables used in the planning model.

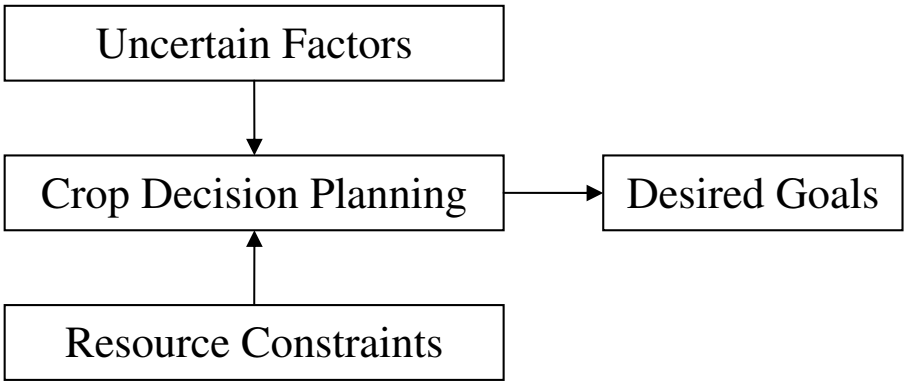

Figure 1: Problems associated in crop decision planning

As yield and price have a tremendous impact on farmers' returns, we need to take into account the uncertainty of these factors to optimize decision planning. Throughout this dissertation, a yield forecasting model under a semiparametric framework is developed using yield history and associated weather data. Next, we outline the price forecasting model based on futures-based approach. Finally, the stochastic crop decision planning model is proposed.

The content of the dissertation is divided into five chapters. We provide a background of the general problem and approaches in Chapter 1. The yield forecasting model is developed in Chapter 2. Chapter 3 presents the price forecasting model. Chapter 4 outlines the crop decision planning model. Conclusions and suggestions of future research are given in Chapter 5.

1.1 Overview of the U.S. Agricultural Industry

Agriculture is a large industry in the U.S. In 2005, agriculture had a value added of approximately \$123.1B, a 25% increase from \$98B in 2000 (Bureau of Economic Analysis 2006). According to Agricultural Resources and Environmental Indicators report (Economic Research Service 2006), there were 21 principal crops that accounted for 95% of harvested crop acreage in the U.S. in 2002. Moreover, only four of these

crops, namely soybean, corn, hay, and wheat, accounted for 80% of all harvested crop acreage. The rest of the harvested acreage was used for fruits, nuts, vegetables, and other minor crops.

Based on the Agricultural Statistics report (National Agricultural Statistics Service 2005), in 2004, *soybean was the most harvested acreage crop among the four major crops*, about 73.9 million acres yielding 3.14 billion bushels and valued 12.2 billion dollars. Most of them were for domestic use and 884 million bushels were exported in 2003. *The second largest crop was corn.* Its harvested area was 73.6 million acres producing 11.8 billion bushels valued at $23B. Similar to soybean, majority of corn crops were designated for domestic use. The third place in the harvested area was hay. It was harvested for 61.9 billion acres with 157.8 million tons of production. Wheat came last in terms of harvest acreage, about 50 million acres. Domestic use and exports were not much different for wheat. *At the state level, Illinois is the first ranked in growing soybean and the second ranked in corn.* Conversely, Iowa is the first ranked in corn and the second ranked in soybean.

In 2004, 2.1 million workers were employed in agriculture, forestry, and fishing (Bureau of Labor Statistics 2005). About 1 million of these workers are self-employed and unpaid family workers, while 1.1 million were wage and salary workers. Almost 50% of the later group was in crop production and 35% was in animal production. The rest of the wage labors was in logging, fishing, forestry, and support activities.

1.2 Overview of Yield Forecasting

Looking closely into the agricultural activities, we find that crop growth and development are largely affected by environmental conditions. These conditions result in significant variation in crop yields from year-to-year and location-to-location. Consequently, understanding the stochastic behavior of crop yield is an essential part at all levels. At the country level, yield forecasting is used in the determination of national

food security, crop insurance policy, import and export plans, and government aid for farmers. At the farm level, knowledge of the yield forecast before the harvest time gives the producers information to plan their farming activities and marketing strategies for their products. For example, predicting shortfalls of crop yields for the coming year gives the government time to initiate appropriate policies and the farmers to make crop selection decisions and to undertake marketing schemes. Consequently, yield forecasting plays an important role in strategic planning and decision making. Crop yield forecasting can be performed via mathematical or statistical methods. Some of the standard methods already investigated include:

Regression analysis - Regression analysis is one of the most widely used methods in yield forecasting (Horie et al. 1992), and many regression models and techniques have been developed (De la Rosa et al. 1981, Garcia-Paredes et al. 2000, Huda et al. 1976, Oberle and Keeney 1990). This technique predicts the response variable, i.e. yield, in terms of explanatory variables such as weather, soil properties, input, and technology. In most of the yield forecasting literature, parametric regression model are used with the assumption that the functional form of the predictor variables is known (Kaspar et al. 2003). The common used models are linear regression models (Shibayama 1991), polynomial regression models (Wilcox et al. 2000), and nonlinear regression models (House 1979).

Simulation - Simulation is often used in yield forecasting. It makes use of meteorological variables such as temperature, rainfall, solar radiation, and humidity to simulate their impact on any agricultural process by a set of mathematical equations, based on the knowledge or experiments of that process. Crop simulation model may be thought as a mathematical representation of the integration of the disciplines of biology, physics, and chemistry (Hoogenboom 2000).

Time series analysis - Time series techniques are often used to analyze crop yield. Some techniques rely exclusively on past yield data. The dependent variable (i.e.

yield) is modeled as a function of time, i.e.

$$Y_t = f(t) + \varepsilon_t,$$

where Y_t is the yield in year t, $f(t)$ is the function that establishes the relationship between yield and time, and ε_t is the error in year t. If the statistical properties of the time series are constant over time, the series is referred to as stationary series. This series can be predicted by simple moving average, simple exponential moving average, autoregressive moving average (ARMA), or autoregressive integrated moving average (ARIMA). On the other hand, if for example, the variance is a function of time, the series is then called nonstationary. Some of the techniques that can handle this kind of data include trend analysis, double moving average, and double exponential smoothing.

1.3 Overview of Price Forecasting

In addition to yield, crop price is another important variable that determines the success of the agricultural business. Crop prices are usually unstable due to the demand and supply of products which highly depend on weather, disease, and pests. The passage of the 1996 Farm Act maintained the market orientation and increased the planting flexibility. Moreover, recent passage of the 2002 Farm Act had an effect on the crop segment through the acreage and production changes which are reflected in the changes of equilibrium levels of prices and demand. As a result, price forecasts are crucial to everyone in the agricultural business from the growers who make production and marketing decisions, to agribusiness companies that buy and sell food products, and to policymakers who manage the commodity programs. Many studies (Adam et al. 1996, Antonovitz and Roe 1986, Byerlee and Anderson 1982, Roe and Antonovitz 1985) address the importance of price forecasting.

For the reasons stated above, several models have been developed to forecast the future cash prices of crops and livestock. A commonly used model is a futures-based

model that relies on the assumption that futures price is a good measure of the actual price (Eales et al. 1990). However, futures price may not reflect the actual local cash price since it represents the world view over the price of the agricultural commodity. Thus, it should be used only as a benchmark. The futures price can be localized by using a commodity basis. Commodity basis is the difference between the local cash price and the relevant futures contract price for a specific time period. It is defined as follows

$$\text{Commodity Basis} = \text{Cash Price - Futures Price.}$$

From the relationship between the local cash price and the futures price through the commodity basis, the expected cash price can be found by the following simple formula

$$\mathbb{E}(\text{Cash Price}) = \text{Futures Price} + \mathbb{E}(\text{Commodity Basis}),$$

where $\mathbb{E}$ denotes the expectation operator.

Regression modeling is also used in price forecasting (Kastens et al. 1998, Kenyon and Kingsley 1973). The predictor variables can be ending stock, production, loan rate, export from other countries, etc. Other researchers use time series analysis techniques in forecasting prices (Liew et al. 2003, Tomek 2000). Crop prices, like other commodity prices, are clearly seasonal. They tend to drop at harvest due to the large amount of crops released to the market and are likely to increase after the harvest time. These time series techniques help reflect the seasonal effect on price and make the forecasting more accurate.

1.4 Overview of Crop Decision Planning

Crop planning involves several decisions including crop and variety selection, acreage allocation, planting, harvesting, storing, and selling. Before each cropping period begins, farmers have to consider which crops they will grow in the coming year. This can be a difficult decision since at the beginning of the cropping season, they do

not have information for weather, yield, price, demand, and supply. They may use their experience or other tools such as advisory service, government crop reports, or computer software, to make decision. Once they decide which crops to grow, they determine the variety of each crop that yields the best return.

Following crop selection, the acreage allocation decision is another issue to consider, particularly when several crops will be produced. Because of the limitation on the land and other resources, the growers should allocate the area among the crops efficiently so that the expected return or utility is maximized.

When the planting period arrives, the cultivation schedule has to be set up according to the planting dates of each crop and the allocation of resources including labor and equipment. The crop fields require close attention since there are many factors that influence the growth of the crops. Flood and drought have a direct impact on crop yield. Weeds and pests also restrain the growth of the produce. Fertilizer, chemicals, and pesticide may be used to enhance the yield.

Similar to planting, the harvesting schedule is constrained by the crops' harvesting dates and the availability of resources. Once the crops are harvested, the next concern is the storage decision. If the crops are storable, farmers may keep them for sale at higher prices after the harvest time. By doing so, producers may obtain higher returns, even after accounting for the storage costs that will occur.

Getting the highest possible return is one of the growers' aims. Selling the products can be made at anytime, even before growing. The easiest selling decision is to sell at harvest. However, this practice may not give a high return because of the seasonal effect. Nonetheless, there are many ways to market the products such as using futures or futures options to hedge the products.

In this dissertation, we develop a crop decision planning model that helps farmers make appropriate decisions under yield and price uncertainties. We forecast crop

yield from historical weather and GDP data under a semiparametric regression framework. The weather variables used in our forecasting model are temperature and rainfall during the growing season. We predict cash price from historical cash and futures prices under a futures-based framework. We incorporate the predicted yield and price from the developed forecasting models in our decision planning model which enables us to explore the effects of stochastic behavior of these uncertainties. The connections between the forecasting models and the decision planning model are graphically illustrated in Figure 2.

The contribution of this dissertation is three-fold: yield forecasting, price forecasting, and crop decision planning. For yield forecasting, we develop a semiparametric regression model that incorporates the within- and between-year relationships in the data. For price forecasting, we propose a functional model-based price forecasting model which estimates the distribution of the commodity basis. For crop decision planning, we develop detailed planning models under stochastic programming and heuristic frameworks.

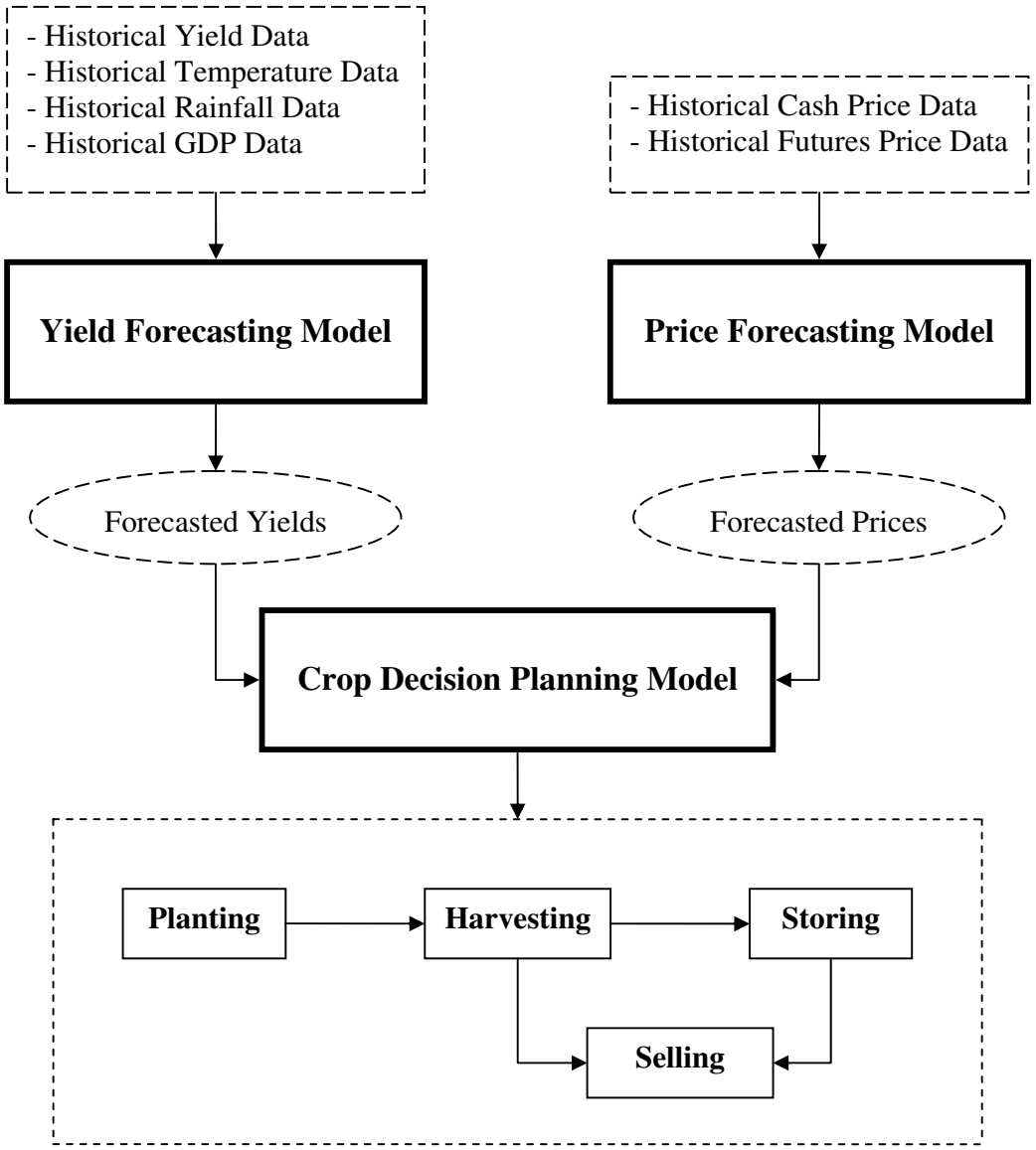

Figure 2: Connections between the crop decision planning model, the yield forecasting model, and the price forecasting model

CHAPTER 2

YIELD FORECASTING

2.1 Introduction

Crop producers often suffer from a lack of accurate information on which to base decisions for crop production and evaluation such as crop yield. Being able to accurately predict yield would allow producers to better prepare for the growing season. *The objective of this chapter is to develop an accurate yield forecasting model along with a prediction confidence band.* This will allow us to explore the effects of yield behavior in the context of a decision planning model.

We develop a crop-weather regression model to forecast the crop yield. In contrast to other regression models that typically use a parametric framework, our proposed yield forecasting model is developed using a semiparametric approach. Within our approach, we incorporate the within- and between-year relationships intrinsic to the data in the yield forecasting model. This results in higher prediction accuracy since we borrow information across the sample data, and therefore, we make better use of the information content in the data. Indeed, *our experimental study shows that the forecasting performance in terms of mean squared error is improved under the semiparametric framework. Another important contribution of this research is the estimation of prediction confidence bands.* The estimated confidence bands are further integrated in the decision planning model.

The layout of this chapter is as follows. The literature is reviewed in Section 2.2. The semiparametric yield forecasting model is described in Section 2.3. The proposed forecasting models is evaluated in Section 2.4. We compare the semiparametric model to a series of other forecasting models for corn and soybean yields and weather in

Illinois. These data are discussed in Section 2.4.1. Fit and prediction results of the yield forecasting models are shown in Section 2.4.5. Conclusions are given in Section 2.5.

2.2 Literature Review

Crop yield forecasting has been a topic of interest for producers, consultants, and agricultural related organizations (Silveira de Jasa 1986). Timely and accurate crop yield forecasts are essential for crop production, marketing, storage, and transportation decisions and they help managing the risk associated with these activities (Bannayan and Crout 1999, Lee 1999, Potgieter et al. 2005). The most well-known and widely used crop information comes from the monthly Crop Production reports (Krog 1988). These reports, prepared by the National Agricultural Statistics Service (NASS), provide statistics and related information of crop production in the U.S. NASS uses data collected from farm operations and field observations to make yield forecasts. The information regarding crop production is provided monthly. The farmers' planting plan is reported in March, while the actual planted acreage is released in June. The monthly yield and production predictions for crop planting for corn and soybean are available in August. At the end of the harvest season, the estimated actual production is also provided (Vogel and Bange 1999). *Even though these reports supply crop yield forecasts that are broadly utilized, they generate only the mean estimate for each state (Lee 1999). This numerical estimate may not reflect the true yield in any particular area in the state.* Moreover, the projected grain yields, which are released monthly from August through November, are not available at the time when farmers need to make decision about planting and production.

According to Walker (1989), there are two distinct crop models - simulation and regression. The strengths and weaknesses of both methods are mentioned by Silveira

de Jasa (1986). A simulation model characterizes the mathematical relationships intrinsic to the data set from previous experiments. This method can generate results under various conditions assuming extensive information used to develop and test the model. However, in agricultural data, in information is rather sparse and incomplete. Because of this limitation, the regression approach is the common approach for predicting yield across large area. In addition, as mentioned by Walker (1989), the same crops are usually cultivated on the same land by the same growers under similar environmental conditions. Therefore, the past yield data contains useful information that can be further used to forecast the current yield production. However, multicollinearity among the predictor variables in multiple regression can be a problem when we want to estimate the contribution of individual predictor.

Various models have been proposed to describe the relationship between yields and related explanatory variables such as weather, soil, water, atmospheric conditions (Ballal et al. 2005, Freckleton et al. 1999, Greenwald et al. 2006, Stephens et al. 1994). The prediction errors associated with the crop models are discussed in detail (Swaney et al. 1986). Horie et al. (1992) give an overview of crop models.

The effects of weather on the crop yields have received wide attention for many years (Guise 1969). Baier (1979) defines the crop-weather models as "a simplified representation of the complex relationships between weather or climate on the one hand and crop performance (such as growth, yield, or yield components), on the other hand by using established mathematical and/or statistical techniques." Many studies incorporate the temperature effect in crop yield prediction (Peng et al. 2004, Wheeler et al. 2000). Sheehy et al. (2006) use the temperature and rice data from 1992 to 2003 with two distinct models, mechanistic model and empirical model, to predict the yield. A mechanistic model is a model based on the underlying physics and chemistry governing the behavior of the process. It uses the knowledge of the interactions between variables to define the model structure. Therefore, it does not

require large data for model development. On the other hand, an empirical model is a data-driven model that specifies the relationship between variables so it heavily depends on data availability. They conclude that the grain yield will decrease by six percent from the based yield for each degree Celsius increased. We find similar temperature effects in other research (Batts et al. 1997, Mitchell et al. 1993, Porter and Gawith 1999).

Rainfall is another explanatory variable that is highly correlated with the yield (Mkhabela et al. 2005, Seif and Pederson 1978). Lomas and Herrera (1985) examine the associates between rainfall and yield of rice grown in Costa Rica from 1975 to 1982. They find that the quadratic regression model gives better prediction results than the simple linear model. August rainfall has the highest correlation, which can explain 52-66% of the variability in rice yield.

Many yield forecasters use several weather variables in their forecasting models. Hoogenboom (2000) provides a comprehensive overview of simulation models using weather variables including temperature, rainfall, and solar radiation. Kandiannan et al. (2002) develop a multiple regression model using rainfall, temperature, evaporation, wind speed, and humidity as the independent variables to predict turmeric yield in Tamil Nadu, India. Data from 1979 to 1999 are used in this analysis. The first 10-year data is used to construct the model and the remaining 10-year data is used to test the model. Even though the coefficient of determination is high, $R^2 = 0.89$, the forecasted values are much different from the observed ones. These differences result in high root mean squared error (RMSE) which is 1,082.7 kg per hectare. They conclude that the non-weather factors, especially technology, may have a considerable influence on the yield.

Beside simulation and regression, there are several weather-based approaches that are applied to crop yield prediction, i.e. Markov chain modeling (Mantis et al. 1985, Mantis et al. 1989) and artificial neural networks (Kaul et al. 2005, Jiang et al.

2004, Park et al. 2005). Markov chain approach is based on the Markov property assumption that the conditional probability distribution of any future state of the process given the past states and the present state is independent of the past states and depends only on the present state. This approach does not have assumptions on random errors like a regression approach. However, Markov chain requires the estimation of transition probability distribution. Artificial neural network (ANN), on the other hand, is based on the human brain's biological neural processes. ANN learns to recognize the patterns or relationships in the data by observing a large number of input and output examples. Once the neural network has been trained, it can predict by detecting similar patterns in future data. Therefore, ANN does not have to specify relationships between dependent and predictor variables in advance. Nevertheless, ANN delivers the results without the explanation of how the results are derived.

Past yield data may be used to estimate the future outcomes without covariate information. For example, Boken (2000) forecasts the spring wheat yields in 1994, 1995, and 1996 by using the yield data sets from 1975-1993, 1975-1994, and 1975-1995, respectively. He applies six time series techniques (linear trend, quadratic trend, simple exponential smoothing, double exponential smoothing, simple moving average, and double moving average) to these data sets and concludes that the quadratic trend performs better than the moving average and the exponential smoothing in terms of mean squared error. These time series analysis techniques, compared to other methods, are easier to implement and use less information. They assume that the surrounding conditions are the same as in past periods and do not take into account the information in the independent variables that may relate to the yield. As a result, the predicted yields may be inaccurate and should be used with caution.

This research focuses on building a crop-weather model using a regression approach. The commonly used approach to yield forecast is linear regression. However,

as studied in this dissertation, there may be nonlinear relationships between yield and weather-based variables. One alternative to allow for nonlinear relationships is to use different degrees of polynomials. One difficulty is to identify the polynomial degree for each predictor where the number of predictors can be very high.

In line with the current literature, we develop a crop-weather model using a regression approach where the weather factors are rainfall and temperature. In contrast to the current approaches, we study a semiparametric model to automatically estimate the nonlinear relationships between yield and weather-based predictors. Additionally, we account for the economic growth by including GDP as a regressor.

2.3 Method

In this section, a semiparametric crop-weather regression model is developed to predict the crop yield. The limitations associated with this model are discussed and the method to surmount these limitations are provided.

2.3.1 Model Formulation

The weather factors used in the proposed model are rainfall and temperature during the growing season. In addition to these factors, we also incorporate the GDP macroeconomic indicator to account for the economic growth, which indirectly reflects in the yield change over time. Since we expect a considerable advance in agricultural technology (i.e. new equipment, better seed, better fertilizer, etc.) over the past few decades, we allow for the technology change through the mean function, which may vary with time only. A functional linear regression analysis is applied to find the relationship between the response variable, which is yield, and the predictor variables, which are temperature, rainfall, and GDP. The initial model incorporates both the weather variables and GDP as follows

$$Y_i = \mu(t_i) + \alpha_1^{(T)}(t_i)T(s_1, t_i) + \ldots + \alpha_m^{(T)}(t_i)T(s_m, t_i) + \alpha_1^{(R)}(t_i)R(s_1, t_i) + \ldots +$$

$$\alpha_m^{(R)}(t_i)R(s_m,t_i)+\alpha^{(GDP)}(t_i)GDP(t_{i-1})+\epsilon_i,\quad i=1,\ldots,N. \tag{1}$$

We assume the errors are identically normally distributed with $\mathbb{E}(\epsilon_i)=0$. In this model, Y_i is the yield observation of the i^{th} year, N is the number of years, and m is the number of months during the growing season. The set of temperature variables consists of $T(s_1,t_i)$, the first month temperature of the growing season in i^{th} year, $T(s_2,t_i)$, the second month temperature of the growing season in i^{th} year, and so on. The set of rainfall variables consists of $R(s_1,t_i)$, the first month rainfall of the growing season in i^{th} year, $R(s_2,t_i)$, the second month rainfall of the growing season in i^{th} year, and so on. We also use the one-year lag nominal GDP - $GDP(t_{i-1})$. We estimate the relationship between the set of predictors above and the yield response by allowing for time (year) dependence in the regression coefficients.

There are several difficulties associated with the yield forecast using the weather-based regression model in (1). First, *it includes a large number of predictors* from GDP and the monthly data of temperature and rainfall. The number of predictors can be even larger when using weekly or daily data and/or other weather predictors are considered (e.g. humidity). Second, *there is a within-year dependence among the predictors*; temperature and rainfall are observed over the growing season within each year and used to predict yearly yields. Consequently, the predictors are not uncorrelated as commonly assumed in regression analysis. Predictor dependence or multicollinearity may not affect the goodness of fit, but the estimated regression coefficients may be unstable due to their joint effect (non-identifiability). Moreover, when forecasting using weekly or daily values for temperature and rainfall, the number of predictors and the correlation between predictors increase dramatically.

One way to overcome these two model limitations is to use principal component analysis (PCA) to transform possibly correlated variables into a smaller set of uncorrelated variables without a significant loss of information. PCA can be found in a wide

range of applications such as computer vision and pattern recognition, source separation, denoising, and biomedical problems (Popovici and Thiran 2004, Raychaudhuri et al. 2000, De la Torre and Black 2001, Schölkopf et al. 1999). PCA makes use of an eigenvalue decomposition of the variance matrix of the data to find the rotating directions and show maximum variabilities on the axes. The eigenvector can be regarded as a weight vector that gives the direction of variability of the corresponding principal component. The eigenvector with the highest eigenvalue determines the direction of the first principal component. This component explains the largest amount of variation in the data. The eigenvector associated with the second largest eigenvalue gives the direction of the second principal component. This component explains the next largest amount of variation and is orthogonal to the first principal component.

The explanatory variables, temperature and rainfall, depend continuously over time and are therefore naturally described as functionals. In order to transform the temperature and rainfall variables into a set of uncorrelated variables and to allow for within-year dependence, we use the functional version of PCA (FPCA). We apply FPCA to temperature and rainfall data separately even though we may expect some degree of collinearity between temperature and rainfall. FPCA applied to functional data from bivariate or multivariate random functions is a research topic that has not yet been explored and it requires rigorous considerations. This topic is beyond to scope of this dissertation. The key references for FPCA are Chapter 8 of Ramsay and Silverman (2005) and Chapter 2 of Ramsay and Silverman (2002), but recently, other methods for estimating FPC's have been introduced. For example, Yao, Müller and Wang (2005) developed a method that allows for a sparse design.

Denote $\omega_j(s,t)$ the weight functions or principal components for temperature data where s is the month index ($s = 1, \ldots, m$) and t is the year index ($t = 1, \ldots, N$) for $j = 1, \ldots, m$. They are functional eigenvectors of the covariance matrix of the

temperature data and they form an orthonormal basis in the sense that

$$\int_s \omega_j^2(s,t)\,ds = 1 \text{ and } \int_s \omega_j(s,t)\omega_k(s,t)\,ds = 0. \tag{2}$$

The temperature scores are defined by

$$P_j(t) = \int_s T(s,t)\omega_j(s,t)\,ds,$$

and have mean zero and $\mathbb{E}(P_j(t)^2) = \lambda_j$, where λ_j is the j^{th} eigenvalue of the temperature covariance matrix. Similarly, for rainfall data, we denote $\rho_j(s,t)$ be the principal components and denote $S_j(t)$ the scores defined by

$$S_j(t) = \int_s R(s,t)\rho_j(s,t)\,ds,$$

where $R(s,t)$ is the rainfall in year t varying with s over the growing season. The rainfall scores also have mean zero and $\mathbb{E}(S_j(t)^2) = \nu_j$, where ν_j is the j^{th} eigenvalue of the rainfall covariance matrix.

The constraints in equation (2) guarantee that the principal components or the weight functions are mutually orthogonal and hence the scores become uncorrelated. Because the first few principal components generally explain most of the variability in the observations, we can reduce the number of variables by discarding the principal components of lesser significance or variability. Let I_T and I_R denote the number of temperature and rainfall principal components selected according to their variability. Further, we replace the temperature and rainfall predictors with the scores corresponding to the first I_T and I_R principal components for temperature and, respectively, for rainfall data. The model becomes

$$Y_i = \mu(t_i) + \alpha_1^{(T)}(t_i)P_1(t_i) + \ldots + \alpha_{I_T}^{(T)}(t_i)P_{I_T}(t_i) + \alpha_1^{(R)}(t_i)S_1(t_i) + \ldots +$$

$$\alpha_{I_R}^{(R)}(t_i)S_{I_R}(t_i) + \alpha^{(GDP)}(t_i)GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{3}$$

We allow for between-year relationships through the regression coefficients, which are functions of time (year).

2.3.2 Model Estimation

We estimate the regression coefficient functions μ, $\alpha_j^{(T)}$ for $j = 1, \ldots, I_T$, $\alpha_j^{(R)}$ for $j = 1, \ldots, I_R$, and $\alpha^{(GDP)}$ using p-splines. We closely follow the estimation procedure of a penalized spline using Best Linear Unbiased Prediction (BLUP) in a mixed model outlined by Ruppert et al. (2003), p. 108-110.

In (3), we assume the following decomposition of the coefficient functions

$$\mu(t) = \beta_0, \quad \alpha_j(t) = \beta_j t + \sum_{k=1}^{K} u_{jk} \left|t - \kappa_k\right|^3 .$$

The decomposition in $\alpha_j(t)$ is based on the radial basis spline functions where κ_k are the fixed knots and K is the number of knots. In the decomposition defined above, β_j are fixed effects and u_{jk} are random effects. We assume that the random effects, u_{jk}, have a normal distribution:

$$\Omega^{1/2} u_{jk} \sim N\left(0, \sigma_u^2 I_K\right), \text{ where } \Omega = \left[\left|\kappa_k - \kappa_{k'}\right|^3\right].$$

Other assumptions in the model are

$$\begin{aligned}
\mathbb{E}\left[P_j(t)\right] = 0 \quad &\text{and} \quad \mathbb{V}\left[P_j(t)\right] = \sigma_j^2 \quad \forall j = 1, \ldots, I_T, \\
\mathbb{E}\left[S_j(t)\right] = 0 \quad &\text{and} \quad \mathbb{V}\left[S_j(t)\right] = \sigma_j^2 \quad \forall j = 1, \ldots, I_R, \\
\mathbb{E}\left[GDP(t)\right] = 0 \quad &\text{and} \quad \mathbb{V}\left[GDP(t)\right] = \sigma^2.
\end{aligned}$$

In addition, $P_j(t)$ and $S_j(t)$ are uncorrelated. These assumptions hold since $P_j(t)$ and $S_j(t)$ are the scores of functional principal components for temperature and rainfall. GDP used in this model is standardized.

Let N be the number of years and define the X and Z matrices as

$$\begin{aligned}
X &= \left[1 \;\; t_i P_1(t_i) \;\; \ldots \;\; t_i P_{I_T}(t_i) \;\; t_i S_1(t_i) \;\; \ldots \;\; t_i S_{I_R}(t_i) \;\; t_i GDP(t_{i-1})\right]_{i=1,\ldots,N}, \\
Z_{T_j} &= \left\{\left[\left|t_i - \kappa_1\right|^3 \;\; \ldots \;\; \left|t_i - \kappa_K\right|^3\right] T_j(t_i)\right\}_{i=1,\ldots,N}, \quad j = 1, \ldots, I_T, \\
Z_{R_j} &= \left\{\left[\left|t_i - \kappa_1\right|^3 \;\; \ldots \;\; \left|t_i - \kappa_K\right|^3\right] R_j(t_i)\right\}_{i=1,\ldots,N}, \quad j = 1, \ldots, I_R, \\
Z_{GDP} &= \left\{\left[\left|t_i - \kappa_1\right|^3 \;\; \ldots \;\; \left|t_i - \kappa_K\right|^3\right] GDP(t_{i-1})\right\}_{i=1,\ldots,N}.
\end{aligned}$$

In order to allow for uncorrelated random effects, we scale Z matrices by $\Omega^{-1/2}$:

$$\tilde{Z}_{T_j} = \Omega^{-1/2} Z_{T_j}, \quad j = 1, \ldots, I_T, \ \tilde{Z}_{R_j} = \Omega^{-1/2} Z_{R_j}, \quad j = 1, \ldots, I_R, \ \tilde{Z}_{GDP} = \Omega^{-1/2} Z_{GDP}.$$

Finally, define $\tilde{Z}$ matrix as

$$\tilde{Z} = \left[\tilde{Z}_{T_1} \ \ \ldots \ \ \tilde{Z}_{T_{I_T}} \ \ \tilde{Z}_{R_1} \ \ \ldots \ \ \tilde{Z}_{R_{I_R}} \ \ \tilde{Z}_{GDP} \right].$$

With the above formulation, (3) can be written in the form of a linear mixed model as

$$Y_i = \beta X + u\tilde{Z} + \varepsilon_i, \quad i = 1, \ldots, N, \tag{4}$$

where β and u are vectors of coefficients.

The approximate 100(1-α)% pointwise prediction band with bias allowance, as adapted from the confidence band proposed by Ruppert et al. (2003), p.137-140, is defined as

$$\hat{Y}(t^*) \pm z_{\left(1-\frac{\alpha}{2}\right)} \hat{\sigma}_\varepsilon \sqrt{C_{t^*} \left(C^T C + \frac{\sigma_\varepsilon^2}{\sigma_u^2} D \right) C_{t^*}^T + 1}, \tag{5}$$

where t^* is the predicted year, $C = \left[1 \ X \ \tilde{Z} \right]$, C_{t^*} is the predicted row of C for $t = t^*$, and $D = diag(0, \ldots, 0, 1, \ldots, 1)$. The number of zeros in D is equal to the number of columns in X plus one (for the intercept) and the number of ones in D is equal to the number of columns in $\tilde{Z}$. Here, we use a pointwise prediction band because we are only interested in the prediction interval of the predicted year.

2.4 Numerical Study

In this section, we provide the numerical study of the semiparametric regression model formulated in Section 2.3. First, we provide the background of the data used in this study. Next, we present the results from FPCA. Then, we analyze the proposed regression model and the linear regression models. Performance of the forecasting models is evaluated and finally, the yield prediction confidence band is estimated.

2.4.1 Data Background

We base our yield forecast on historical corn and soybean yield data, weather data, and nominal GDP. Historical corn and soybean yield data are acquired from Quick Stats, an agricultural statistics database, provided by the National Agricultural Statistics Service. Yield data are expressed as a number of bushels harvested per acre. Both corn and soybean yield data are from Hancock County in Illinois from 1927 to 2005. We chose Illinois as our primary state in our study since based on the Agricultural Statistics report (National Agricultural Statistics Service 2005), Illinois is the largest soybean producer, and the second largest corn producer in the U.S. Hancock county is chosen as a representative county in Illinois because its corn and soybean yields in 2005 are about the same level as the state average. However, our methodology applies to any crop producer across the country.

In our weather-based model, we use the weather data from National Climatic Data Center (NCDC). These data are collected from La Harpe station in Hancock County, Illinois, from 1927 to 2005. The rainfall variable is the total monthly rainfall in inches and the temperature variable is the average monthly temperature in degree Fahrenheit during the cropping season. Based on the Usual Planting and Harvesting Dates for US Field Crops report (National Agricultural Statistics Service 1997), in Illinois, corn is usually planted around the end of April and harvested in late September. Similarly, soybean is planted in the beginning of May and harvested in late September. Therefore, we include only the temperature and rainfall data from May to September in our study.

Another variable used in yield forecasting is annual GDP from 1926 to 2004. We use the nominal GDP acquired from Economic History Services. We utilize the one-year lag nominal GDP since the current yield is the reflection of past year's economic growth.

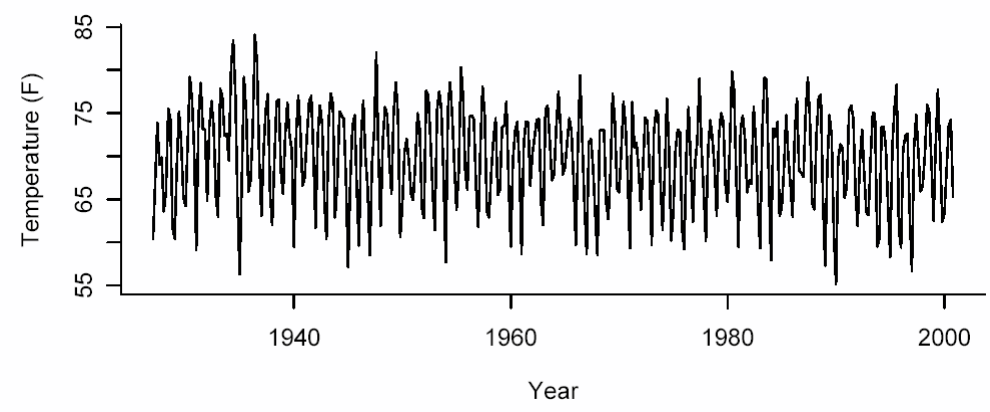

Figure 3: Time series of temperature from May to September (1927 to 2000)

2.4.2 Functional Principal Component Analysis

We apply the functional principal component analysis (FPCA) described in Section 2.3.1 to temperature and rainfall data. The temperature data explored in this study is the average monthly temperature in degrees Fahrenheit and rainfall data is the total monthly rainfall in inches from May to September. We forecast one-year ahead or one-lag. This means we use data from 1927 to 1995 to forecast the yield for 1996, data from 1927 to 1996 to forecast the yield for 1997, and so on. For demonstration purpose, we provide the FPCA results only for the data from 1927 to 2000.

The time series plots of temperature and rainfall data are depicted in Figures 3 and 4, respectively. We observe the seasonal pattern in temperature time series but not in rainfall.

The smoothed mean function of the temperature data during the growing season is shown in Figure 5. We find that the average temperature gradually increases from May and reaches the highest average temperature at 76°F in July, indicated by the dashed line, and then steadily decreases until the end of the harvesting period in September. This pattern is as expected since the temperature is highest during the summer period and lower in spring and fall.

For the rainfall data, the smoothed mean function is displayed in Figure 6. The

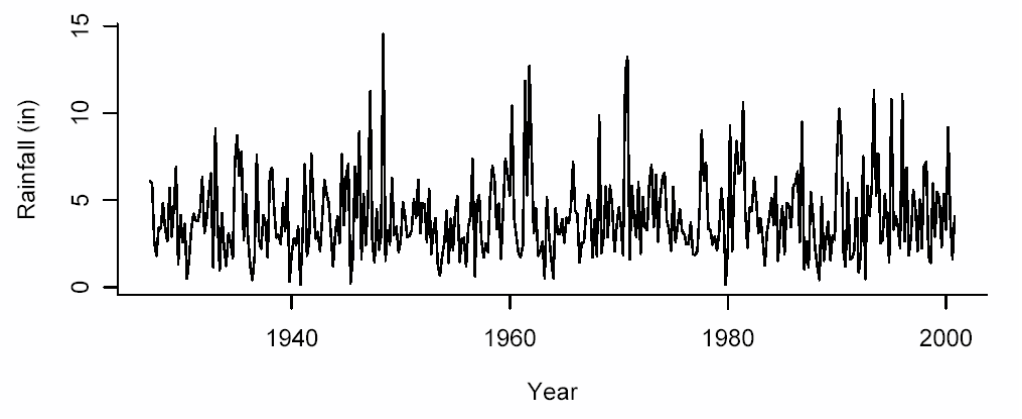

Figure 4: Time series of rainfall from May to September (1927 to 2000)

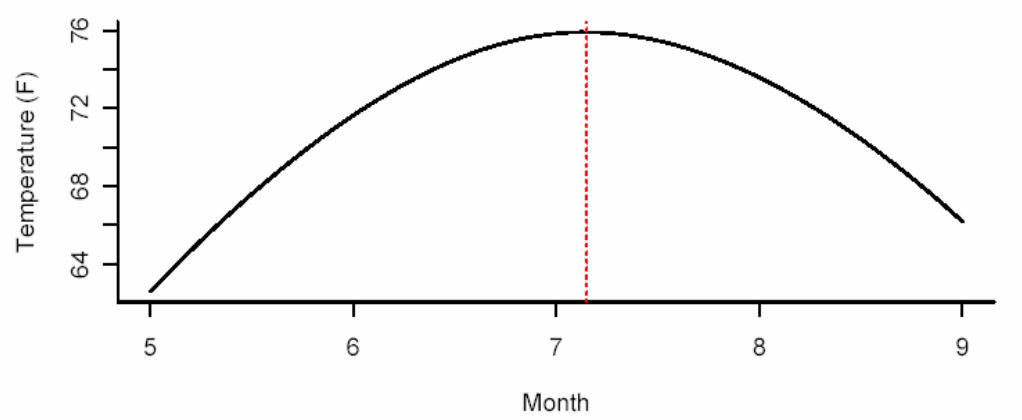

Figure 5: Smoothed mean function of temperature from May to September (1927 to 2000)

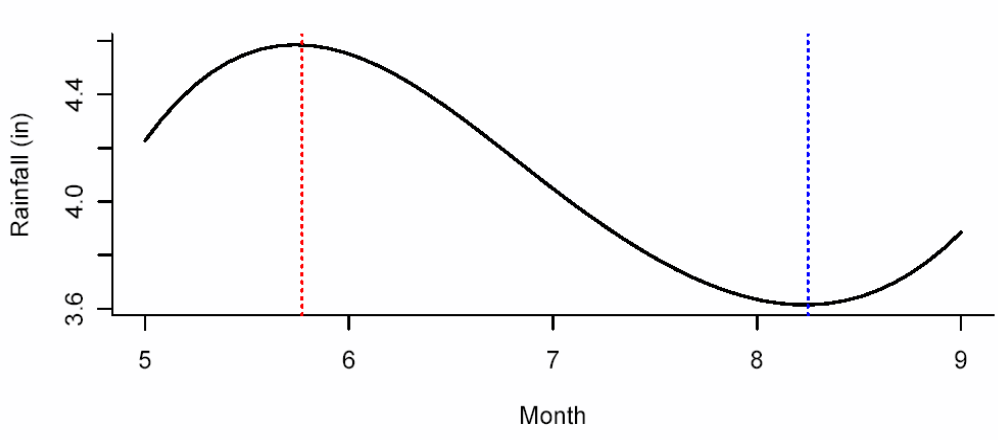

Figure 6: Smoothed mean function of rainfall from May to September (1927 to 2000)

total rainfall is highest in the beginning of June, indicated by the left dashed line, and decreases rapidly until it reaches its minimum in August, indicated by the right dashed line.

Next, we compute the principal component weight function, which explain the amount of variation in a decreasing order. The number of principal components equals the number of time points. Since there are five months, May to September, there will be five principal components. Figure 7 depicts all principal component curves for the temperature data. Each panel shows the weight function for the temperature data after the mean across 73 years has been removed from each month. The first weight function, displayed in the upper left panel, is negative throughout the year. The highest absolute weight is placed on July (recall that it is the month that has the highest temperature). The lowest absolute weight is assigned to September temperature, which is about a half of the highest absolute weight. This implies that the greatest variability between years can be found by heavily weighting May to August and lightly weighting September. The second weight function is displayed in the upper right panel. This function has a sinusoidal shape. The weight gradually

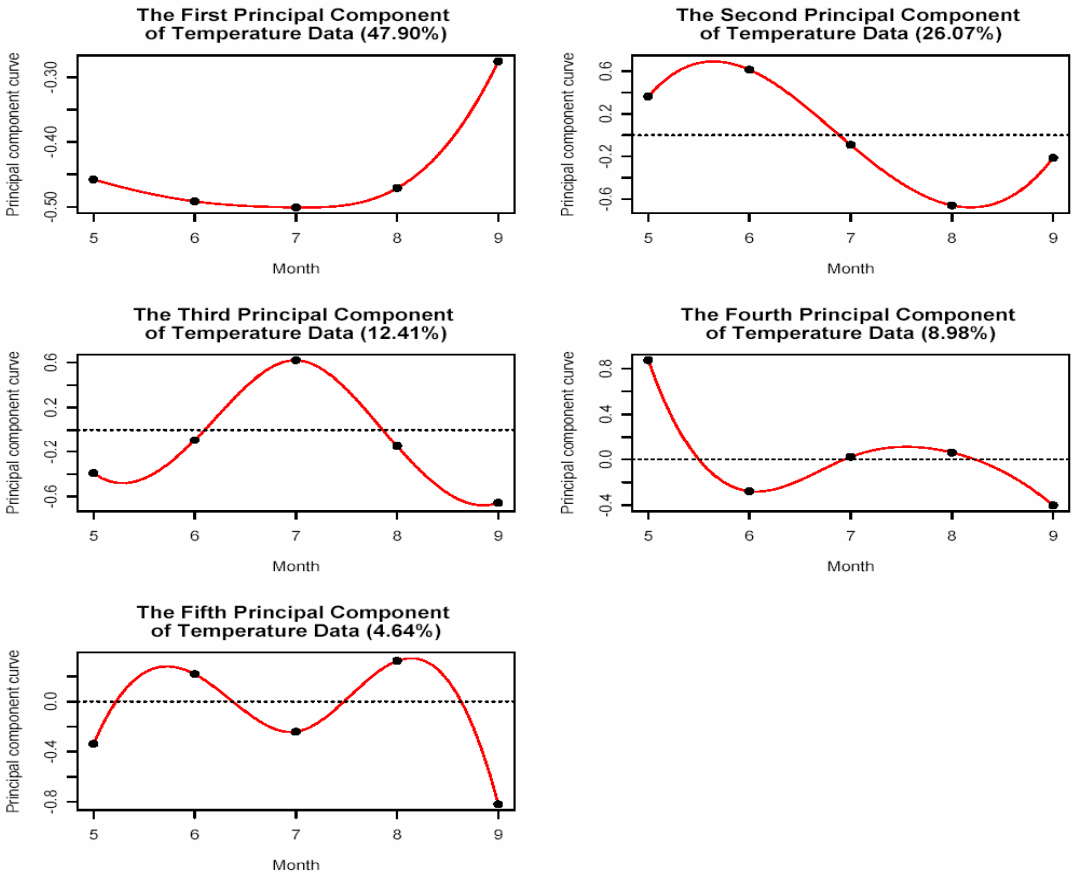

Figure 7: Principal component curves of temperature data from 1927 to 2000

increases from May until it reaches its peak in June. After that, it decreases passing zero in July to the minimum in August and then increases again in September. This component consists of a positive contribution for temperature before July and a negative contribution for temperature after July.

The principal component curves of the rainfall data are displayed in Figure 8. Similar to the temperature data, each panel shows the weight function after subtracting the mean over all 73 years from each monthly rainfall data. The first weight function is shown in the upper left panel. The highest absolute weight is placed on July and moderate absolute weight on June. Small positive weights are assigned to August and September and a small negative weight is assigned to May. In contrast, the second weight function, displayed in the upper right panel, places positive weights

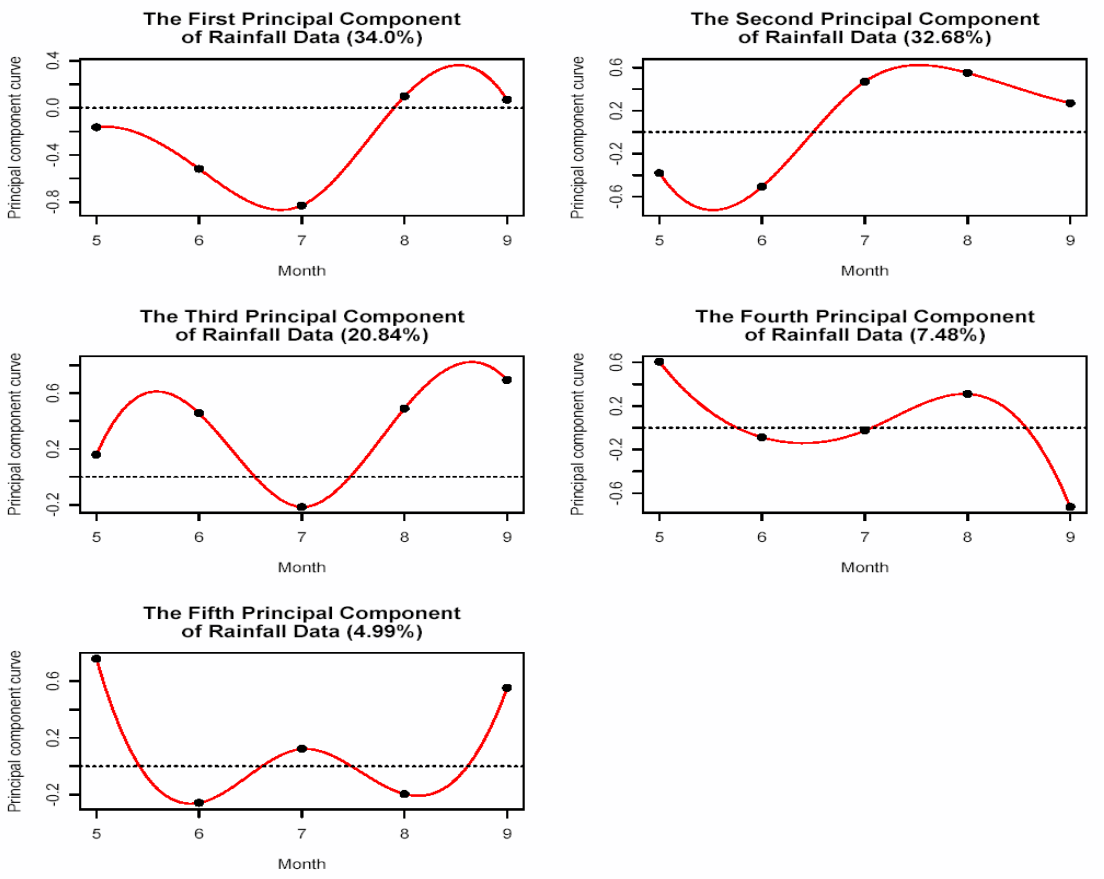

Figure 8: Principal component curves of rainfall data from 1927 to 2000

on July, August, and September and negative weights on May and June. Thus, this component has a negative effect from the first two months and a positive effect from the last three months.

Figures 9 and 10 show the proportion of variance explained by each principal component of temperature and rainfall data, respectively. The first temperature principal component accounts for 47.90% of total variation while the second accounts for 26.07%. These two principal components account for almost three-fourths of the variability. Therefore, we may use only the first two principal components of the temperature data in the yield forecasting model. By incorporating only the first two principal components, we reduce the number of parameters in the yield forecasting model by $3 \times K_T$ where K_T is the number of non-zero coefficients used for estimating

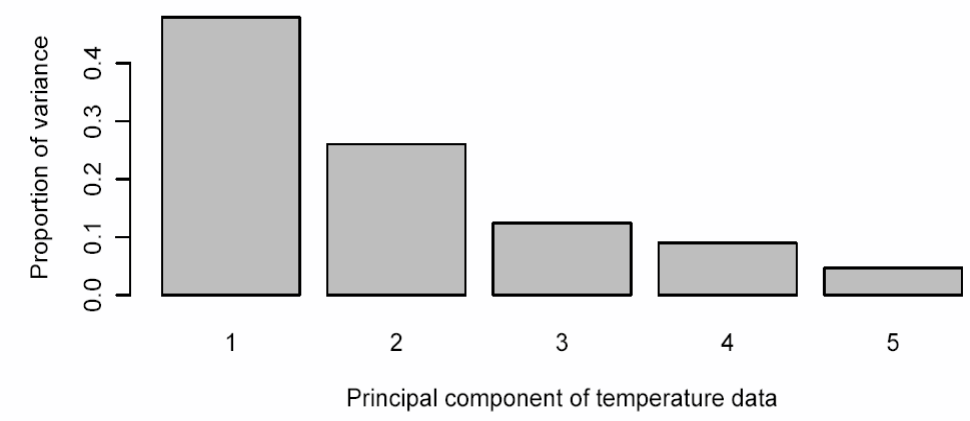

Figure 9: Bar plot of the variance proportions explained by the five principal components of temperature data from 1927 to 2000

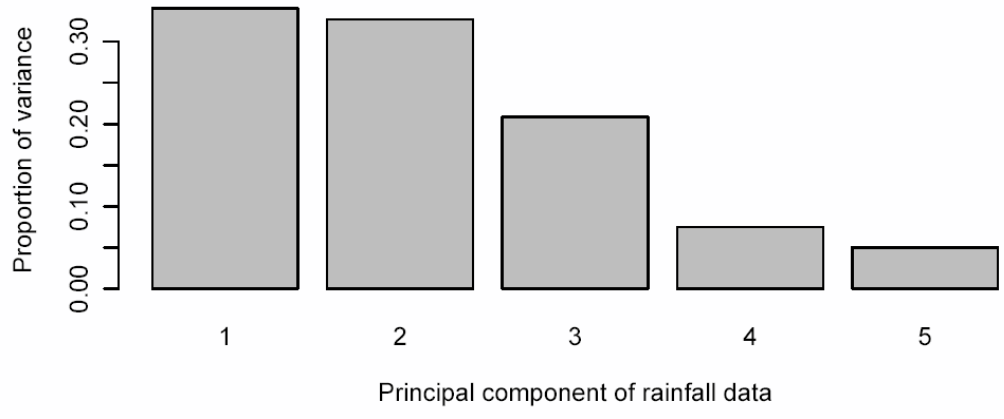

Figure 10: Bar plot of the variance proportions explained by the five principal components of rainfall data from 1927 to 2000

the linear functional $\alpha^{(T)}(t)$ in equation (3).

Likewise, the first rainfall principal component explains 34% of the variability. The second component accounts for 32.68% and the third component accounts for 20.84% of the variation. Thus, the first three principal components altogether explain 87.52% of the total variability. Consequently, by using the first three principal components, the number of coefficients in the yield forecasting model reduces from $5 \times K_R$ to $3 \times K_R$ where K_R is the number of non-zero coefficients of $\alpha^{(R)}(t)$, the coefficient functional of a rainfall principal component.

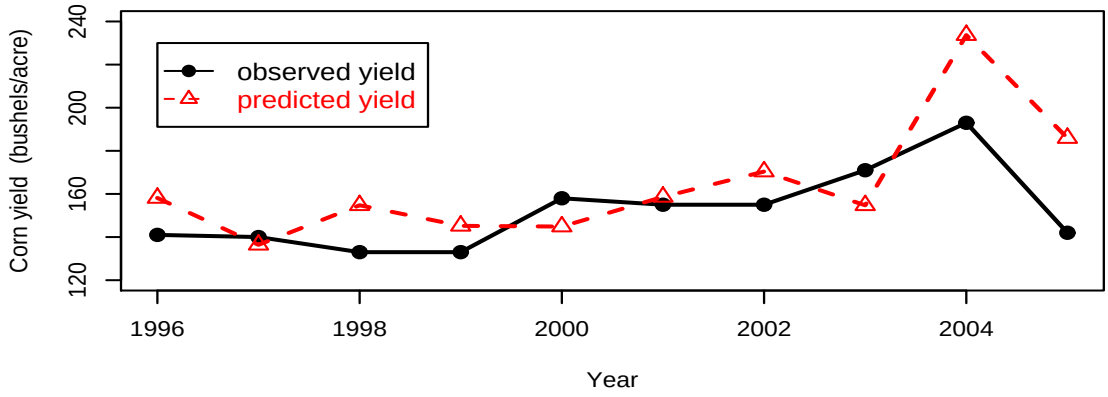

Figure 11: Observed corn yield and predicted corn yield as provided by Model 1

2.4.3 Additive Regression Model

2.4.3.1 Corn

In corn yield forecasting, the scores of temperature and rainfall principal components are used as predictors in the semiparametric regression model as described in (3). First, we incorporate the scores of all principal components along with the standardized lag nominal GDP. The model as defined in (3) for $I_T = 5$ and $I_R = 5$ can be written as

$$Y_i = \mu(t_i) + \alpha_1^{(T)}(t_i)P_1(t_i) + \ldots + \alpha_5^{(T)}(t_i)P_5(t_i) + \alpha_1^{(R)}(t_i)S_1(t_i) + \ldots +$$

$$\alpha_5^{(R)}(t_i)S_5(t_i) + \alpha^{(GDP)}(t_i)GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{6}$$

We refer this model as *Model 1*. The one-lag prediction results for 1996 to 2005 are illustrated in Figure 11. This model provides good predictions for only a few years with large prediction errors for the beginning and the end of the prediction period.

According to the amount of variation explained by each principal component score described in Section 2.4.2, the first two temperature principal component scores and

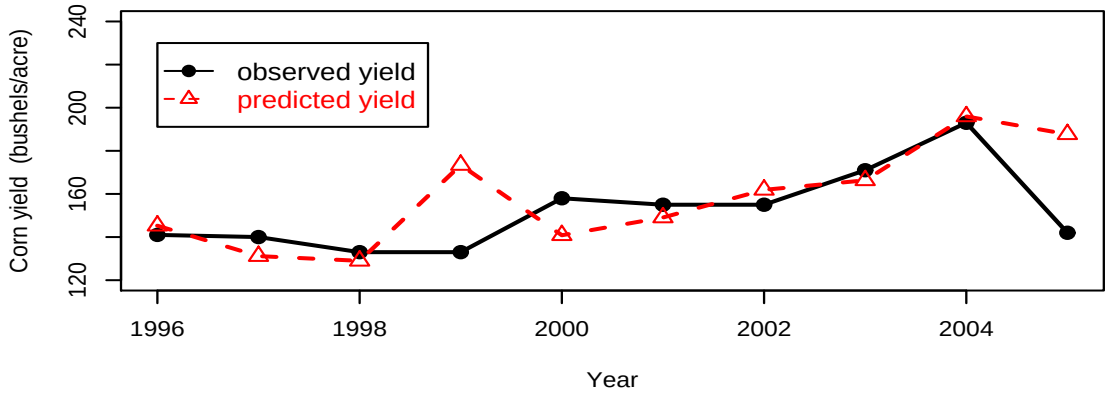

Figure 12: Observed corn yield and predicted corn yield as provided by Model 2

the first three rainfall principal component scores explain most of the variation in the data. This suggests using only these five principal component scores and the standardized lag nominal GDP to predict the yearly yield. The model is described in equation (3) for $I_T = 2$ and $I_R = 3$, which we refer to *Model 2*. The model becomes

$$Y_i = \mu(t_i) + \alpha_1^{(T)}(t_i)P_1(t_i) + \alpha_2^{(T)}(t_i)P_2(t_i) + \alpha_1^{(R)}(t_i)S_1(t_i) + \alpha_2^{(R)}(t_i)S_2(t_i) +$$

$$\alpha_3^{(R)}(t_i)S_3(t_i) + \alpha^{(GDP)}(t_i)GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{7}$$

Figure 12 depicts the one-year ahead forecasting results for 1996 to 2005. This model delivers a better forecast than the previous model even though it has large prediction errors in 1999 and 2005. This implies that discarding high order principal components improves the performance of the model.

The output of the semiparametric regression model (not shown here) indicates that some coefficient functions of weather principal component scores have high smoothing parameter values which imply that they are approximately linear. This suggests using linear functions to estimate some of the regression coefficients. Using linear rather than nonlinear coefficient functions entails a more parsimonious model, which will be

easier to predict and interpret. We search exhaustively starting with the full model of corn yield forecasting (Model 1) to identify a set of predictors and the shape of their coefficient functions (linear vs. nonlinear) that will provide the best overall prediction and fitting with respect to one-lag prediction mean squared error. *We use smoothing parameter values as a criterion to determine the shape of the coefficient functions and t-test criterion to select the set of linear predictors.* This gives the final model (*Model 3*). This model consists of two nonparametric nonlinear components, standardized lag nominal GDP and first temperature principal component score, three linear components, second and third temperature principal component scores, and first rainfall principal component score. The final corn yield forecasting model is in the equation below

$$Y_i = \mu(t_i) + \alpha_1^{(T)}(t_i)P_1(t_i) + \alpha_2^{(T)}P_2(t_i) + \alpha_3^{(T)}P_3(t_i) +$$

$$\alpha_1^{(R)}S_1(t_i) + \alpha^{(GDP)}(t_i)GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{8}$$

The one-year ahead predicted yields for 1996 to 2005 are shown in Figure 13. The predicted yields are very close to the observed yields. There is a moderate error in 2005.

2.4.3.2 Soybean

Soybean yield forecasting model also uses the temperature and rainfall principal components and the standardized lag nominal GDP as the regressors. Therefore, when incorporating all principal components in the model, the soybean yield forecasting model will have the same Model 1 as corn yield forecasting model (6). In addition, when using only the major principal component scores, the first two temperature principal component scores and the first three rainfall principal component scores, soybean's Model 2 will also be the same as corn's Model 2 (7). Figures 14 and 15

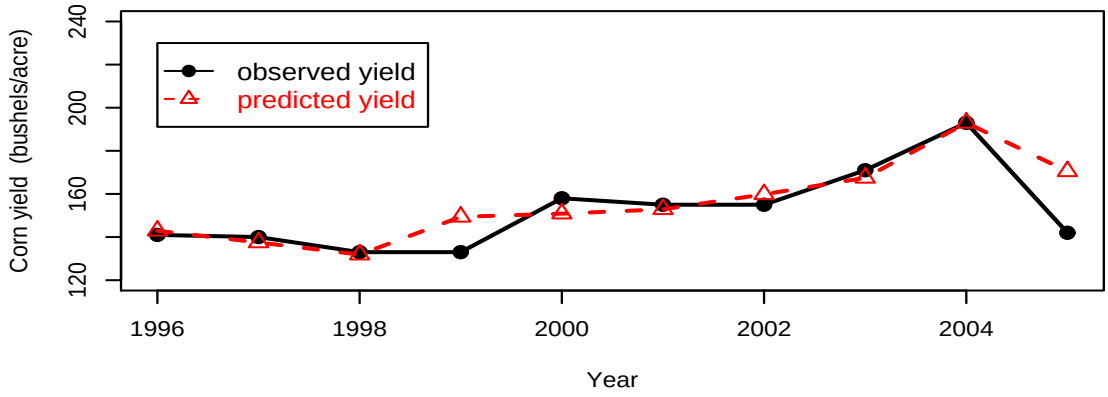

Figure 13: Observed corn yield and predicted corn yield as provided by Model 3

depict the one-lag prediction results for 1996 to 2005 of Models 1 and 2, respectively. Overall, both models deliver good yield predictions. However, Model 1 has large prediction errors in 1999 and 2004 while Model 2 has large prediction discrepancy only in 1999.

We perform an exhaustive search to find the final semiparametric regression model for soybean. This results in a model with two nonlinear components, standardized lag nominal GDP and second rainfall principal component score, and four linear components, first, second, and fifth temperature principal component scores and first rainfall principal component score. This model is referred as *Model 3*:

$$Y_i = \mu(t_i) + \alpha_1^{(T)} P_1(t_i) + \alpha_2^{(T)} P_2(t_i) + \alpha_5^{(T)} P_5(t_i) + \alpha_1^{(R)} S_1(t_i) +$$

$$\alpha_2^{(R)}(t_i) S_2(t_i) + \alpha^{(GDP)}(t_i) GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{9}$$

The one-year ahead forecasting results are illustrated in Figure 16. This model provides better forecasts than the first two models even though it still has a large prediction error in 1999.

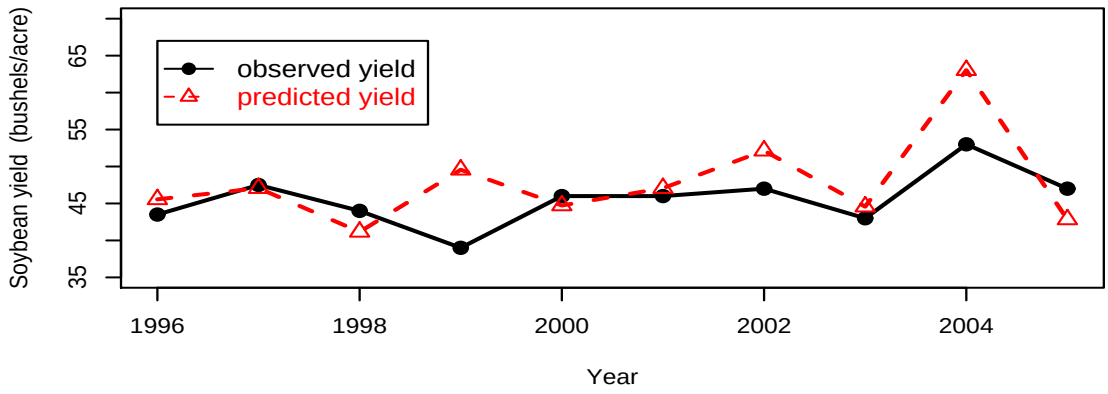

Figure 14: Observed soybean yield and predicted soybean yield as provided by Model 1

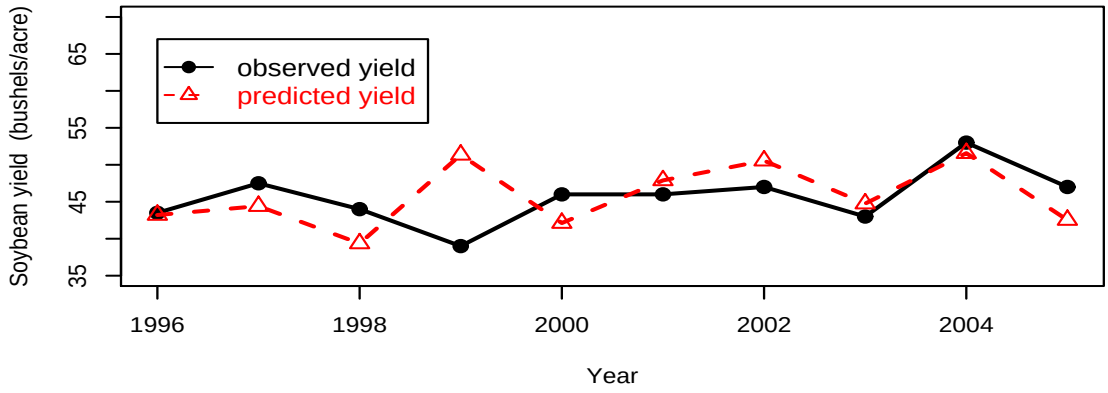

Figure 15: Observed soybean yield and predicted soybean yield as provided by Model 2

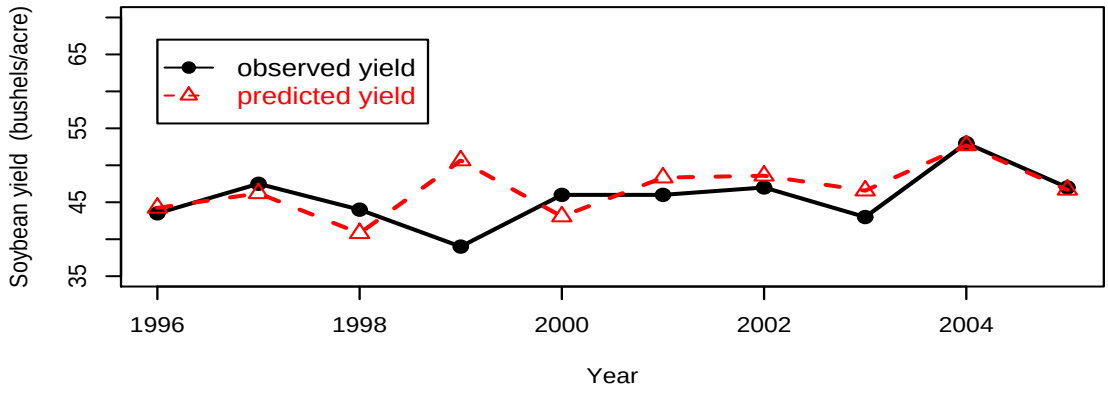

Figure 16: Observed soybean yield and predicted soybean yield as provided by Model 3

2.4.4 Linear Regression Analysis

The common approach to predicting yield from weather data is linear regression (Kandiannan et al. 2002, Sheehy et al. 2006, Seif and Pederson 1978) as provided by the model below

$$Y_i = \mu(t_i) + \alpha_1^{(T)} T(s_1, t_i) + \ldots + \alpha_m^{(T)} T(s_m, t_i) + \alpha_1^{(R)} R(s_1, t_i) + \ldots +$$

$$\alpha_m^{(R)} R(s_m, t_i) + \alpha^{(GDP)} GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{10}$$

The difference between (1) and the linear regression in (10) is that the coefficients of the latter model are fixed over time. We perform the model selection under the linear regression model using t-test criteria to obtain the final linear regression model for corn and soybean. The results are as follows.

2.4.4.1 Corn

Using t-test criteria, we obtain a corn yield forecasting model with five predictor variables including standardized lag nominal GDP, May, July, and August temperature,

and July rainfall. This model is referred to *Model 4* which is given in the equation below

$$Y_i = \mu(t_i) + \alpha^{(T)}_{May}\, T(May, t_i) + \alpha^{(T)}_{Jul}\, T(Jul, t_i) + \alpha^{(T)}_{Aug}\, T(Aug, t_i) +$$

$$\alpha^{(R)}_{Jul}\, R(Jul, t_i) + \alpha^{(GDP)} GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{11}$$

Another version of the linear regression model is obtained by replacing the monthly temperature and rain predictors with the scores corresponding to their functional principal components. The model selection using t-test criterion is applied to the linear regression with the functional principal component scores as predictors resulting in a model with five predictor variables including standardized lag nominal GDP, first three temperature principal component scores, and first rainfall principal component score. This model is referred to *Model 5*:

$$Y_i = \mu(t_i) + \alpha^{(T)}_1 P_1(t_i) + \alpha^{(T)}_2 P_2(t_i) + \alpha^{(T)}_3 P_3(t_i) +$$

$$\alpha^{(R)}_1 S_1(t_i) + \alpha^{(GDP)} GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{12}$$

The one-lag prediction results of Models 4 and 5 for 1996 to 2005 are provided in Figures 17 and 18, respectively. These two linear regression models provide good predictions with moderate errors in years 1999 and 2005 and small errors in the first three years.

2.4.4.2 Soybean

We apply the same model selection procedure as we did in Section 2.4.4.1 to soybean data. The model using monthly weather data, *Model 4*, consists of standardized lag nominal GDP, July and August temperature, and May, July, and August rainfall. This model becomes

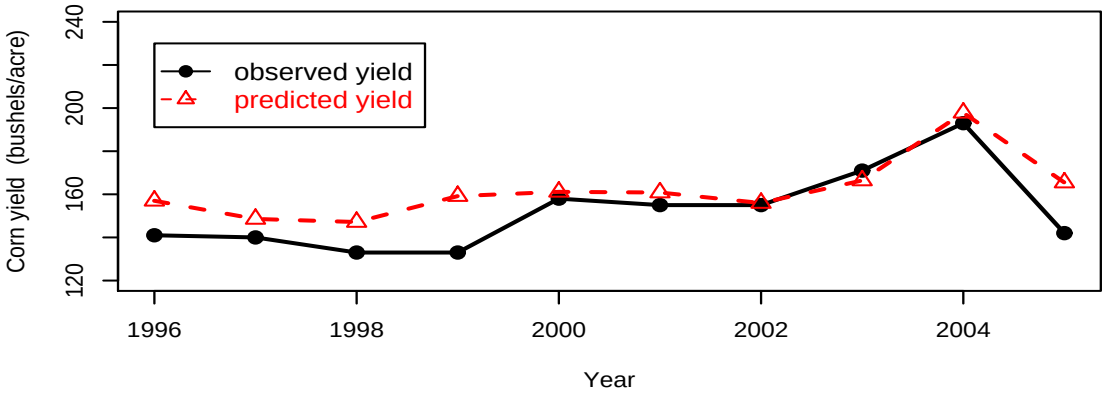

Figure 17: Observed corn yield and predicted corn yield as provided by Model 4

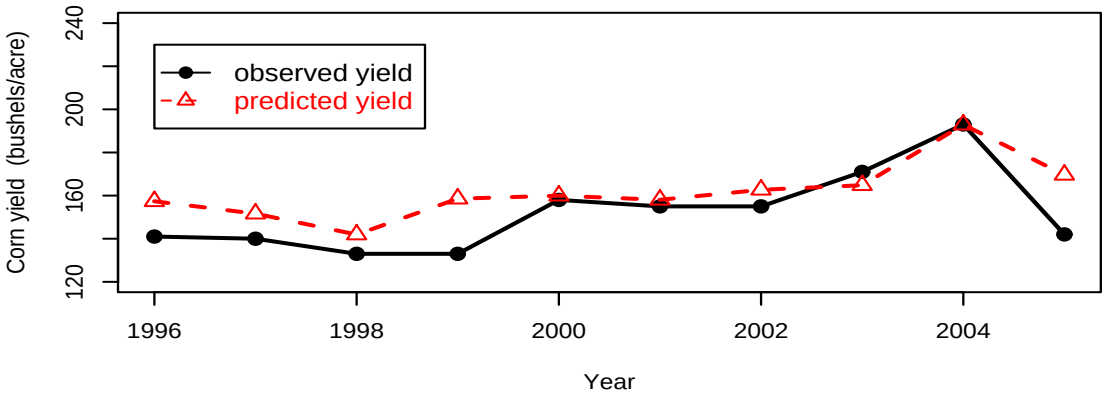

Figure 18: Observed corn yield and predicted corn yield as provided by Model 5

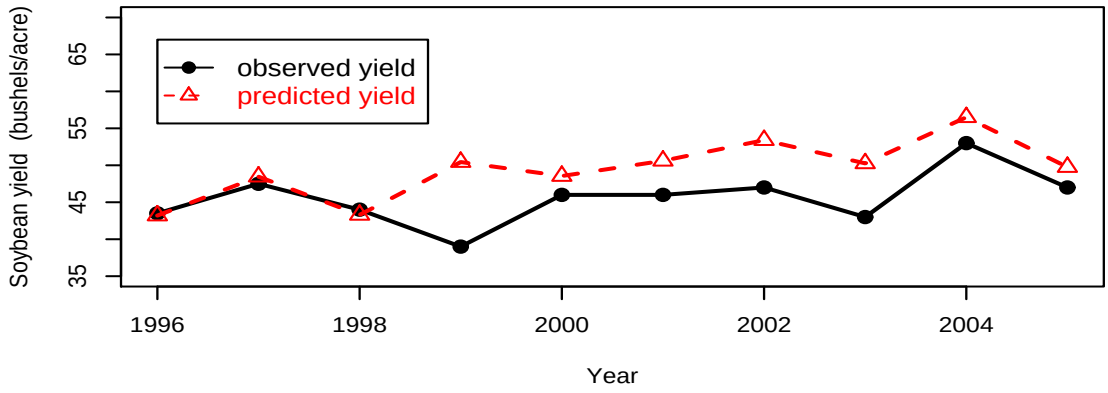

Figure 19: Observed corn yield and predicted soybean yield as provided by Model 4

$$Y_i = \mu(t_i) + \alpha^{(T)}_{Jul}\, T(Jul, t_i) + \alpha^{(T)}_{Aug}\, T(Aug, t_i) + \alpha^{(R)}_{May}\, R(May, t_i) +$$

$$\alpha^{(R)}_{Jul}\, R(Jul, t_i) + \alpha^{(R)}_{Aug}\, R(Aug, t_i) + \alpha^{(GDP)} GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{13}$$

Likewise, the model using principal component scores of the weather data, referred as *Model 5*, consists of standardized lag nominal GDP, first and second temperature principal component scores and second and third rainfall principal component scores:

$$Y_i = \mu(t_i) + \alpha^{(T)}_1 P_1(t_i) + \alpha^{(T)}_2 P_2(t_i) + \alpha^{(R)}_2 S_2(t_i) +$$

$$\alpha^{(R)}_3 S_3(t_i) + \alpha^{(GDP)} GDP(t_{i-1}) + \epsilon_i, \quad i = 1, \ldots, N. \tag{14}$$

Figures 19 and 20 illustrate the one-year ahead soybean yield forecasting results of Models 4 and 5, respectively. These models deliver good predictions in half of the predicted years. There are a large error in 1999 and moderate errors from 2000 to 2004.

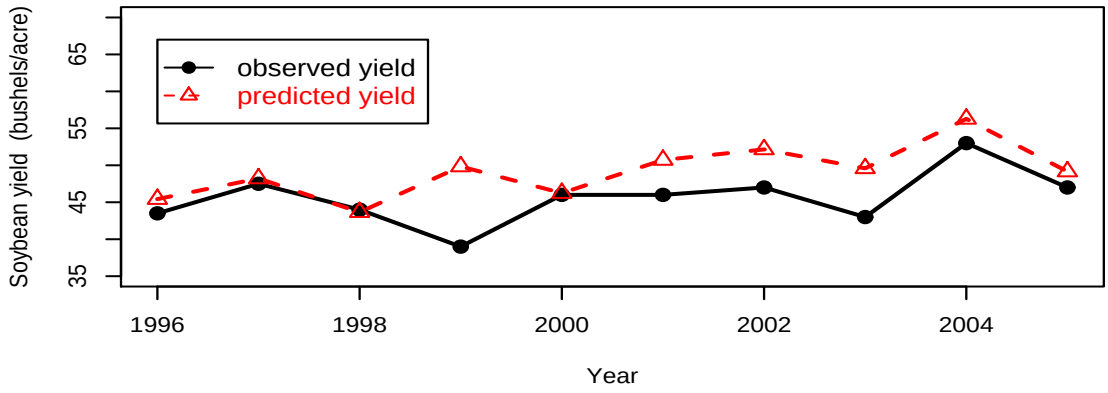

Figure 20: Observed corn yield and predicted soybean yield as provided by Model 5

2.4.5 Model Evaluation

We use the mean squared error (MSE) criterion to evaluate the performance of the forecasting models. Table 1 summarizes the performance of these models for corn yield one-lag prediction as provided by MSE. *Both parametric regression models, Models 4 and 5, have moderate MSE. The semiparametric regression models 1 and 2, on the other hand, have much higher MSE than the linear regression models 4 and 5. The selected model, Model 3, has the smallest MSE.* It is about one-half of MSE for Models 4 and 5, and one-fourth of MSE for semiparametric models 1 and 2. Note that the observed corn yield in year 2005 is lower than the prediction in all five models. This is because there were extreme drought conditions during the 2005 growing season (Zhang et al. 2006). The proposed models can capture part of this drought effect. This can be seen from lower forecasted yield in 2005 than one in 2004.

The performance of soybean yield forecasting models are provided in Table 2. *Model 1 has MSE similar to Model 4. On the other hand, Model 2 has the same MSE level as Model 5. The selected model, Model 3, has the smallest MSE. In addition, it has the highest adjusted coefficient of determination.* From corn and soybean yield

Table 1: Corn yield prediction results from observed weather data for Model 1 to Model 5

Predicted Year	Observed Yield (bushel/acre)	Model 1	Model 2	**Model 3**	Model 4	Model 5
1996	141	158.13	145.27	142.91	157.00	157.37
1997	140	136.42	131.16	137.50	148.58	151.62
1998	133	154.75	128.98	131.86	147.15	141.93
1999	133	145.21	173.43	149.40	159.18	158.57
2000	158	144.87	140.84	150.86	161.15	159.96
2001	155	158.64	149.07	152.89	160.83	158.06
2002	155	170.39	161.90	159.85	155.97	162.70
2003	171	154.74	166.31	167.47	166.42	164.81
2004	193	233.66	195.96	192.99	197.82	192.89
2005	142	185.92	187.68	170.49	165.45	169.63
MSE		519.75	424.15	**118.34**	181.41	201.11
R^2 †		90.96%	89.52%	89.21%	79.30%	78.90%
R^2-adj †		**89.45%**	88.63%	88.47%	77.80%	77.40%

† Using data from 1927 to 2004

forecasting results, *we can conclude that allowing for time-dependent relationships reflected in the nonlinear coefficient functions together with the functional principal component analysis improve the performance of the forecasting model.*

In practice, we do not know the weather condition in advance. Therefore, we also need the weather forecast in the yield forecasting model. For weather forecast, we use a standard time series analysis technique called autoregressive integrated moving average (ARIMA). We forecast for weather one year ahead. Since the model uses only the weather from May to September, we calibrate the forecasted weather of these months by the difference between the means of the first four months of the predicted and observed data in that year. The resulting weather forecasts are close to the true observed values. *We perform a similar model evaluation as for the observed weather and find that Model 3 still has the smallest MSE for both corn and soybean* (see Tables 3 and 4). This is as expected since the same models provide the best yield forecast when the weather values are assumed to be known.

Table 2: Soybean yield prediction results from observed weather data for Model 1 to Model 5

Predicted Year	Observed Yield (bushel/acre)	Model 1	Model 2	**Model 3**	Model 4	Model 5
1996	43.5	45.56	43.21	44.24	43.20	45.43
1997	47.5	47.07	44.41	46.21	48.46	48.22
1998	44	41.16	39.36	40.81	43.29	43.66
1999	39	49.56	51.35	50.65	50.44	49.84
2000	46	44.73	42.12	43.07	48.55	46.24
2001	46	47.07	47.89	48.33	50.61	50.72
2002	47	52.14	50.56	48.59	53.40	52.15
2003	43	44.57	44.75	46.57	50.26	49.59
2004	53	63.03	51.57	52.69	56.50	56.28
2005	47	42.84	42.53	46.68	49.76	49.16
MSE		27.35	24.00	**17.75**	27.38	22.96
R^2 †		87.81%	86.06%	86.87%	77.60%	77.30%
R^2-adj †		85.77%	84.88%	**85.96%**	75.70%	75.80%

† Using data from 1927 to 2004

Table 3: Corn yield prediction results from forecasted weather data for Model 1 to Model 5

Predicted Year	Observed Yield (bushel/acre)	Model 1	Model 2	**Model 3**	Model 4	Model 5
1996	141	144.53	133.89	143.59	160.13	158.19
1997	140	147.92	135.81	139.86	154.79	153.50
1998	133	168.38	138.68	124.33	136.62	135.96
1999	133	150.81	152.41	143.12	151.52	150.68
2000	158	145.33	132.69	128.72	141.74	141.47
2001	155	161.17	154.78	154.38	161.95	161.64
2002	155	166.93	161.62	154.56	158.21	158.40
2003	171	179.32	179.53	177.36	178.61	177.30
2004	193	184.22	173.92	160.45	161.63	160.18
2005	142	174.51	183.87	163.35	163.97	161.98
MSE		318.79	355.16	**259.81**	278.88	264.40

Table 4: Soybean yield prediction results from forecasted weather data for Model 1 to Model 5

Predicted Year	Observed Yield (bushel/acre)	Model 1	Model 2	**Model 3**	Model 4	Model 5
1996	43.5	44.98	43.62	45.70	48.20	47.87
1997	47.5	46.45	43.97	46.25	48.09	47.25
1998	44	45.53	41.69	45.99	45.54	43.82
1999	39	49.18	48.50	50.14	49.74	48.31
2000	46	43.54	40.76	44.62	46.33	45.18
2001	46	49.08	47.34	48.76	50.76	50.05
2002	47	48.15	45.16	47.88	49.69	48.81
2003	43	48.93	47.71	48.90	53.19	52.70
2004	53	45.82	41.36	46.02	48.75	48.48
2005	47	42.25	43.60	47.83	50.47	49.41
MSE		23.54	30.99	**22.90**	30.40	24.66

2.4.6 Prediction Confidence Band

Since Model 3 of corn and soybean provides the best prediction, we use it as the base model to determine the pointwise prediction band defined in Section 2.3.2. Predicted corn yield for year 2005, as shown in Table 1, is 170.49 bushels per acre. The 95% pointwise prediction band for year 2005 in bushels per acre is **(131.32, 209.66)**. The observed corn yield in year 2005 is 142 bushels per acre. It is close to the lower bound of the estimated confidence interval due to the severe drought conditions mentioned in Section 2.4.5. On the other hand, the predicted soybean yield for 2005 is 46.68 bushels per acre. The 95% pointwise prediction band for year 2005 in bushels per acre is **(37.94, 55.42)**.

The pointwise prediction bands of Model 3 are illustrated in Figure 21 for corn and Figure 22 for soybean. The prediction band will give a range of possible outcomes, and will allow us to cope with different scenarios that may occur over the planning horizon for crop decisions.

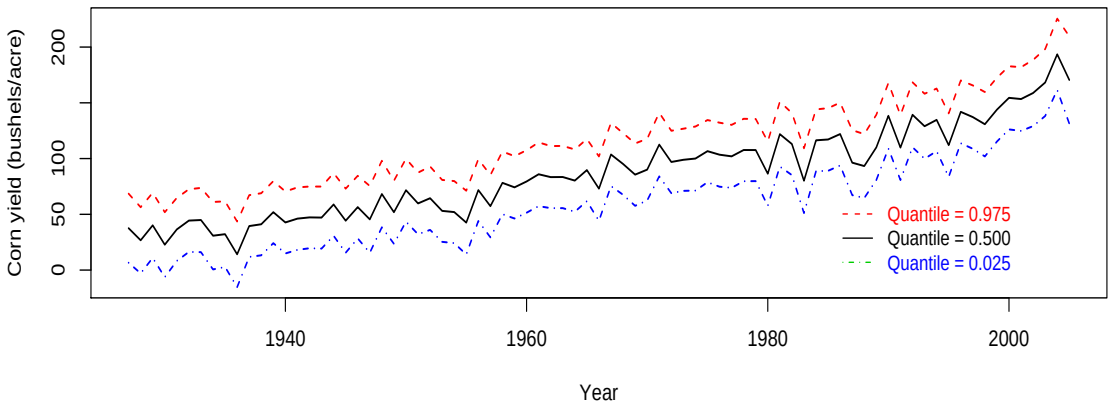

Figure 21: Plot of 95% pointwise prediction confidence band of the fitted corn yield from 1927 to 2005

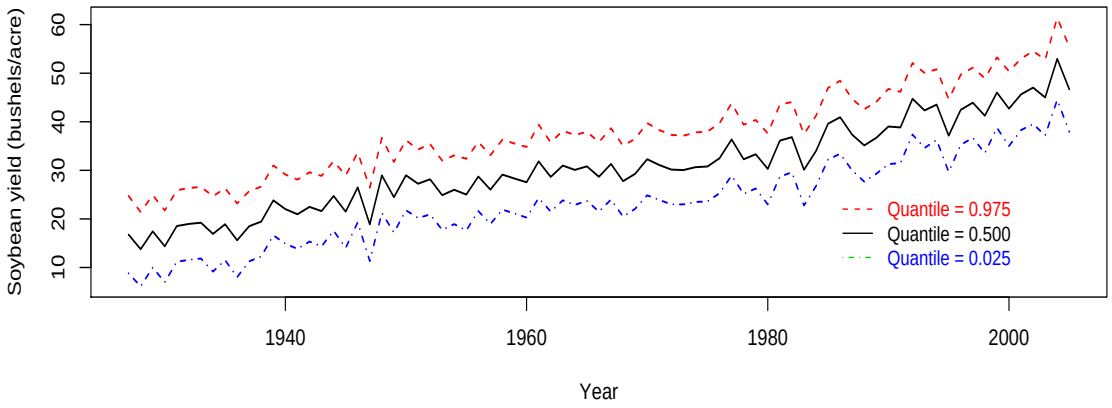

Figure 22: Plot of 95% pointwise prediction confidence band of the fitted soybean yield from 1927 to 2005

2.5 Conclusions

To the best of our knowledge, parametric linear regression has been used as the forecasting tool in almost all studies on yield forecasting that use a regression approach. Linear regression has been mainly considered since it is easy to use and interpret. However, as pointed out in Section 2.3, it does not incorporate the between-year relationships in the weather data. Disregarding these relationships may result in inaccurate prediction as already discussed in Section 2.4.5. Subsequently, we proposed a semiparametric regression model, which allows for these relationships. In addition, we incorporate the concept of principal component analysis in our proposed model. This technique reduces dimensionality of the model without much loss of information. It also transforms the correlated variables into uncorrelated variables.

We develop the final semiparametric regression model for corn and soybean from the full model that incorporates the whole set of functional principal component scores. The selected model is compared with the linear regression models, the full semiparametric model, and the semiparametric model that uses only the major functional principal component scores. Since our yield forecasting models make use of future weather information, in practice, we need to forecast the weather condition. In this research, we employ the time series technique called ARIMA to forecast for weather one year ahead. Finally, the yield prediction confidence band is estimated and will be used in the crop decision planning model in Chapter 4.

According to the numerical study on the data discussed in Section 2.4 , our selected semiparametric model outperforms other models in both corn and soybean yield predictions. It has the smallest mean squared error and high adjusted coefficient of determination when using observed and forecasted weather data. This result confirms that the semiparametric regression model with the functional principal component scores enhances the forecasting performance.

Another observation arises from this forecasting analysis. Temperature gives more

information than rainfall, especially in the corn yield prediction. In the final corn yield prediction model, we use three temperature principal component scores and only one rainfall principal component score. In soybean yield forecasting, we use three temperature principal component scores and only two rainfall principal component scores.

CHAPTER 3

PRICE FORECASTING

3.1 Introduction

The profitability of an agribusiness is heavily influenced by the crop prices since they directly affect costs and revenues. Crop producers need to know the crop prices in order to select the crops that provide the highest return and determine the proper time to sell their products. Because many decisions are made in advance before the prices are realized, price forecasting is crucial for the producers and agribusinesses. *This chapter focuses on developing the price forecasting model that provides accurate forecasts in the timely manner.*

We develop a cash price forecasting model under a futures-based framework, which predicts the cash price from futures contract price and commodity basis. In this research, we concentrate on forecasting commodity basis rather than cash price. We apply the concept of functional model-based clustering analysis to estimate the density function of the commodity basis. Our model is distinct from other futures-based models that usually estimate the expected commodity basis. Hence, *the main contribution of this chapter is to estimate the commodity basis distribution. The distribution is used to estimate the confidence interval of commodity basis and cash price, and will be further integrated in the crop decision planning model.*

This chapter is organized in the following order. Section 3.2 reviews the literature in this area. The methodology used in this research is presented in Section 3.3. Section 3.4 presents the numerical study on corn and soybean price forecasting. The conclusions are given in Section 3.5.

3.2 Literature Review

Estimation of future selling prices of crops is an important piece of information for farmers since it helps determine what crops they should plant and how much return they would receive. Farmers usually plant the crops that can generate the highest profits. To achieve this objective, knowledge of future crop price is necessary. However, since commodity prices are volatile, growers and agribusinesses have to rely on price forecasting. To address their needs, the U.S. Department of Agriculture (USDA) publishes reports on crop supply, demand, yields, prices, and other related information. Agricultural prices are provided by the World Agricultural Outlook Board (WAOB) in the World Agricultural Supply and Demand Estimates (WASDE) report. As mentioned by Isengildina et al. (2004), these prices are interval estimates rather than point estimates, and they are forecasted based on several methods and information sources along with expert judgment. However, the WASDE report is distributed only once a month and the prices are the average national prices which differ from the prices received at a particular location. Forecasted prices are required to be more location specific (Kastens et al. 1998). Therefore, USDA predicted prices should only be used as a benchmark (Irwin et al. 1994, Isengildina et al. 2004).

Kenyon and Lucas (1998) study the relationship between soybean season average prices and soybean ending stocks. The ending stocks are calculated by subtracting total demand from total supply, which become the beginning stocks for the next crop year. They find that prices tend to decrease if the ending stocks increase compared to the beginning stocks, and supply increases compared to demand in each crop year and vice versa. They propose a simple price forecasting model using price historical data and the ending stocks based on linear regression. The key is to estimate demand and supply of each crop. The difference between supply and demand (i.e. the ending stocks) will determine the prices of the crops in that year. A similar approach has been applied to corn, wheat, and cotton.

A number of models have been developed to forecast the cash prices. Many researchers study the role of futures contract prices in agricultural price forecasting (Dow 1940, Gardner 1976, Kenyon et al. 1993, Tomek and Gray 1970, Working 1942). Futures price is often used as an indicator of the expected cash price (Hoffman 2005). Eales et al. (1990) examine the difference between futures prices and the means of the aggregate price distributions surveyed from farmers and grain merchandisers in Illinois. In most cases, futures price and aggregate price are not significantly different so this result suggests that futures prices can be used to estimate the expected price. Just and Rausser (1981) compare the performance of the spot price forecasts among commercial firms (Chase Econometrics, Doanes Agricultural Service, Data Resources, Inc. (DRI), and Wharton Econometric Forecasting Associates), USDA, and futures market. The commodities used in this research are corn, wheat, soybean, soybean oil, soybean meal, cotton, live cattle, and hogs. To compare the forecast performance, root mean squared error and root mean squared percentage error are evaluated. No model performs consistently well over all commodities. However, futures prices do better on the average.

Working (1942) and Tomek and Gray (1970) examine the difference between cash price and futures price. They define this difference as the "cost of carrying" indicating the incentive to hold the stock for later use. This cost can be either positive (reflecting large inventory) or negative (reflecting tight inventory). This cost is called the *commodity basis*.

Generally, in the agribusiness literature, *commodity basis* is defined as the difference between the local cash price and the price of a futures contract for a specific time period. It reflects the local market conditions which are influenced by several factors, such as local supply and demand conditions, interest, storage costs, transportation costs, handling costs, and profit margins.

It is common to analyze the performance of futures-based models and determine a

commodity basis model that provides the best forecast. Kastens et al. (1998) explore the accuracy of three futures-based cash price forecasting models with two simple forecasting models under the economic assumptions of the efficient market hypothesis and the law of one price. Prices for a range of commodities and locations from the first week of 1982 to the last week of 1996 are investigated. Comparing the mean absolute percentage error from each model, they conclude that the deferred futures plus the most recent five-year average commodity basis and the deferred futures plus level and proportional commodity basis perform better than the other three models. They also use the forecast error regression model to explain the forecast errors. Overall, the deferred futures price plus five-year historical average commodity basis performs the best. Hauser et al. (1990) compare the naïve and market-based soybean basis expectation models. They conclude that simpler models give reasonably good soybean basis forecasts.

In this research, a price forecasting model is developed under a futures-based framework where the cash price is forecasted from the futures price and commodity basis. We predict the cash price by obtaining a forecast of the commodity basis distribution over one year. We derive the cash price forecast by adding the commodity basis forecast to the futures price. In order to estimate the one-year commodity basis distribution, we use a functional model-based approach. As we obtain the distribution rather than the expectation alone, we can also compute a confidence band for the one-year commodity basis forecast.

3.3 Method

In this section, we develop a cash price forecasting model under the futures-based framework where cash price is forecasted from futures price and commodity basis. Next, we estimate the distribution of the commodity basis and derive its pointwise confidence band.

3.3.1 Model Formulation

We focus on forecasting the commodity basis rather than the cash price because the commodity basis is less volatile than the cash price, and the futures price can be found from many resources such as futures markets, newspapers, and internet. Therefore, we forecast the cash price by first obtaining a forecast of the commodity basis over one year and then adding the futures price to it. We divide the N-year commodity basis data into N different functional observations, each observation consisting of commodity basis values observed over a one-year period. Consequently, our data are both functional and longitudinal

$$Y_j(t_i), \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, n = N.$$

Our goal is to identify common patterns among the N years and use them to predict the commodity basis of the upcoming year. We adopt a model-based approach to estimate the density function of the commodity basis distribution. Model-based clustering, introduced by Banfield and Raftery (1993), relies on estimation of a mixture density function. Each component in the mixture corresponds to one cluster. Within this method, the main assumption is that the observations $y_1, \ldots, y_n$ are random variables from a mixture distribution with K components. However, in Banfield and Raftery (1993), Celeux and Govaert (1995), and Dasgupta and Raftery (1998), the model framework does not allow for functional relationships of the data. Since our commodity basis data are functional, we instead exploit a functional data model-based framework. Consequently, we follow the approach for clustering functional data proposed by James and Sugar (2003).

We assume that the predicted commodity basis curve belongs to a model component with some probability as estimated using the mixture likelihood approach. Under this approach, the cluster memberships Z_j's are treated as missing data assuming that Z_j for $j = 1, \ldots, n$ are multinomial with parameters $(\pi_1, \ldots, \pi_K)$ and

π_k is the probability that a commodity basis curve belongs to the k^{th} cluster. Let $f_k(y_j|\theta_k)$ be the density function corresponding to the k^{th} cluster, parameterized by θ_k. The parameters are estimated by maximizing

$$L(\theta_1, \ldots, \theta_K; \pi_1, \ldots, \pi_K | y_1, \ldots, y_n) = \prod_{j=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_j|\theta_k). \tag{15}$$

Let b_{ij}, β_{ij}, and ε_{ij} be, respectively, the observed commodity basis value, true commodity basis value, and measurement error in year j and time t_i, i.e. $b_{ij} = b_j(t_i)$ and $\beta_{ij} = \beta_j(t_i)$. The commodity basis model can be formulated as

$$b_{ij} = \beta_{ij} + \varepsilon_{ij}, \quad t_i = t_1, \ldots, t_{n_j}, \quad j = 1, \ldots, n,$$

where n is the number of years and n_j is the number of time points in year j. This model assumes β_{ij} follows a Gaussian process, and the measurement errors have mean zero and are uncorrelated with each other. For each year j, we expand the true function $\beta_j(t)$ using a set of spline basis functions, and for each group k, we compute the mean, $\mu_k(t)$, along with the cluster proportion parameter π_k as extensively discussed in James and Sugar (2003) and summarized in the next section. Under the mixture likelihood framework

$$y(t) \sim \sum_{k=1}^{K} \pi_k f(\mu_k(t), \Sigma_k(t)), \tag{16}$$

where K is the number of clusters, f is the density function of the cluster, and Σ_k is the covariance matrix of the k^{th} cluster.

We apply the functional model-based clustering to the standardized commodity basis data to forecast the commodity basis distribution on a common scale for all years.

3.3.2 Model Estimation

We expand the true commodity basis value by a set of spline basis functions $\beta_j(t) = s(t)^T \varphi_j$, where $s(t)$ is a vector of spline basis and φ is a spline coefficient vector. The

spline coefficients are modeled by assuming a Gaussian distribution:

$$\varphi_j = \mu_{z_j} + \gamma_j, \quad \gamma_j \sim N(0, \Gamma),$$

where μ_{z_j} is the cluster mean or the cluster fixed effect, z_j is the unknown cluster membership of year j, and γ_j is a random effect of year j. Define $S_j = \big(s(t_1), \ldots, s(t_{n_j})\big)^T$ to be the spline basis matrix corresponding to the j^{th} year, b_j to be the vector of the observed values, and ε_j to be the vector of measurement errors. The functional clustering model (FCM) can be written as

$$b_j = S_j(\mu_{z_j} + \gamma_j) + \varepsilon_j, \quad j = 1, \ldots, n, \tag{17}$$

$$\varepsilon_j \sim N(0, \sigma^2 I), \quad \gamma_j \sim N(0, \Gamma).$$

In this model, we assume the covariances of the ε_j's and γ_j's to be, respectively, $\sigma^2 I$ and Γ for all clusters. Therefore, under this formulation, the distribution of β_j is

$$b_j \sim N(S_j\mu_{z_j}, \Sigma_j), \text{ where } \Sigma_j = \sigma^2 I + S_j \Gamma S_j^T. \tag{18}$$

As in FCM, we estimate the parameters by maximizing the mixture likelihood function (15). The estimation procedure is fully described in James and Sugar (2003).

Under the Gaussian process assumption, we assume that our predicted curve follows a mixture of normals whose k^{th} component has mean μ_k and variance Σ_k. We can further derive the $100(1-\alpha)\%$ pointwise confidence band of the functional mixture model by

$$\hat{b}(\alpha, t) = \sum_{k=1}^{K} \hat{\pi}_k \left(\hat{\mu}_k(t) \pm \Phi^{-1}(1 - \frac{\alpha}{2}) \hat{\upsilon}_k(t) \right), \tag{19}$$

where Φ^{-1} is the inverse cumulative density function and υ_k is the diagonal of the covariance matrix of the k^{th} cluster.

3.4 Numerical Study

In this section, we illustrate price forecasting as described in Section 3.3 for corn and soybean. First, we compute commodity basis and investigate the common patterns. Next, we perform the functional clustering analysis on the standardized commodity basis data. Subsequently, we estimate the prediction confidence band and calibrate it to its normal scale. Finally, we determine the predicted cash price from the forecasted commodity basis and futures price.

3.4.1 Data Background

We base our price forecast on both futures and cash price data. They are used to calculate the commodity basis by subtracting futures price from the corresponding cash price. We focus on forecasting the commodity basis because it typically does not vary as much as cash price and can generally be predicted from historical commodity basis patterns (Chicago Board of Trade 2000).

The futures price data are acquired from an agricultural package provided by the Chicago Board of Trade. This package provides a number of futures prices of corn and soybean for every trading day. These prices include Open, Close, High, Low, and Settlement prices. Each day, there are several futures contracts available for trading. Investors can trade several years in advance. A futures contract is classified by its delivery month or contract month. However, there are specific delivery months for each crop. Corn futures contracts are delivered only in March, May, July, September, and December. On the other hand, soybean futures contracts can be delivered in January, March, May, August, September, and November. In this research, we select the *nearby settlement price* to represent the futures price. A *nearby contract* is the futures contract that is closest to expiration. For example, December corn futures is the nearby futures for corn in October. The *settlement price* is determined by averaging a range of closing prices.

We acquire the cash price data from the USDA Springfield regional office. This office provides the average cash prices of corn and soybean traded in central Illinois. Both futures and cash prices are collected every business day from 1991 to 2005. We compute the commodity basis from nearby futures and cash prices.

3.4.2 Commodity Basis Information

We construct the commodity basis history for corn and soybean by subtracting futures price from the corresponding cash price. The commodity basis history spans from 1991 to 2005. We will use the commodity basis history from 1991 to 2004 to explore and predict commodity basis pattern in 2005. Further, we will estimate the cash price by adding the forecasted commodity basis to the expected futures price.

3.4.2.1 Corn Basis

The corn basis plots from 1991 to 2005 are displayed in Figure 23. From these plots, we can observe similar basis patterns across years. Clear similarity is for adjacent years, i.e. corn basis in 1991-1992, 1998-1999, 2000-2002, and 2003-2005. Overall, the basis fluctuates during the year with five common local maxima and one local minimum. This pattern is clearly displayed in 2001 and 2002 plots. In these plots, the first four maxima are indicated by four dashed lines starting from the left of the plots. The lowest basis is underlined with a straight line while the last local maximum is marked by the dashed line on the right of the plots. The 1996 corn basis is an outlier since it behaves differently from other years. This outlier may come from the passage of the 1996 Farm Act, which increased the planting flexibility, and resulted in a large amount of corn released to the market.

Table 5 shows the associated dates and corn basis values at the local extrema. The first local maximum usually occurs at the end of February while the second occurs at the end of April. The third local extremum, on the other hand, varies from the end of June to the middle of July. The fourth local extremum takes place at the end of

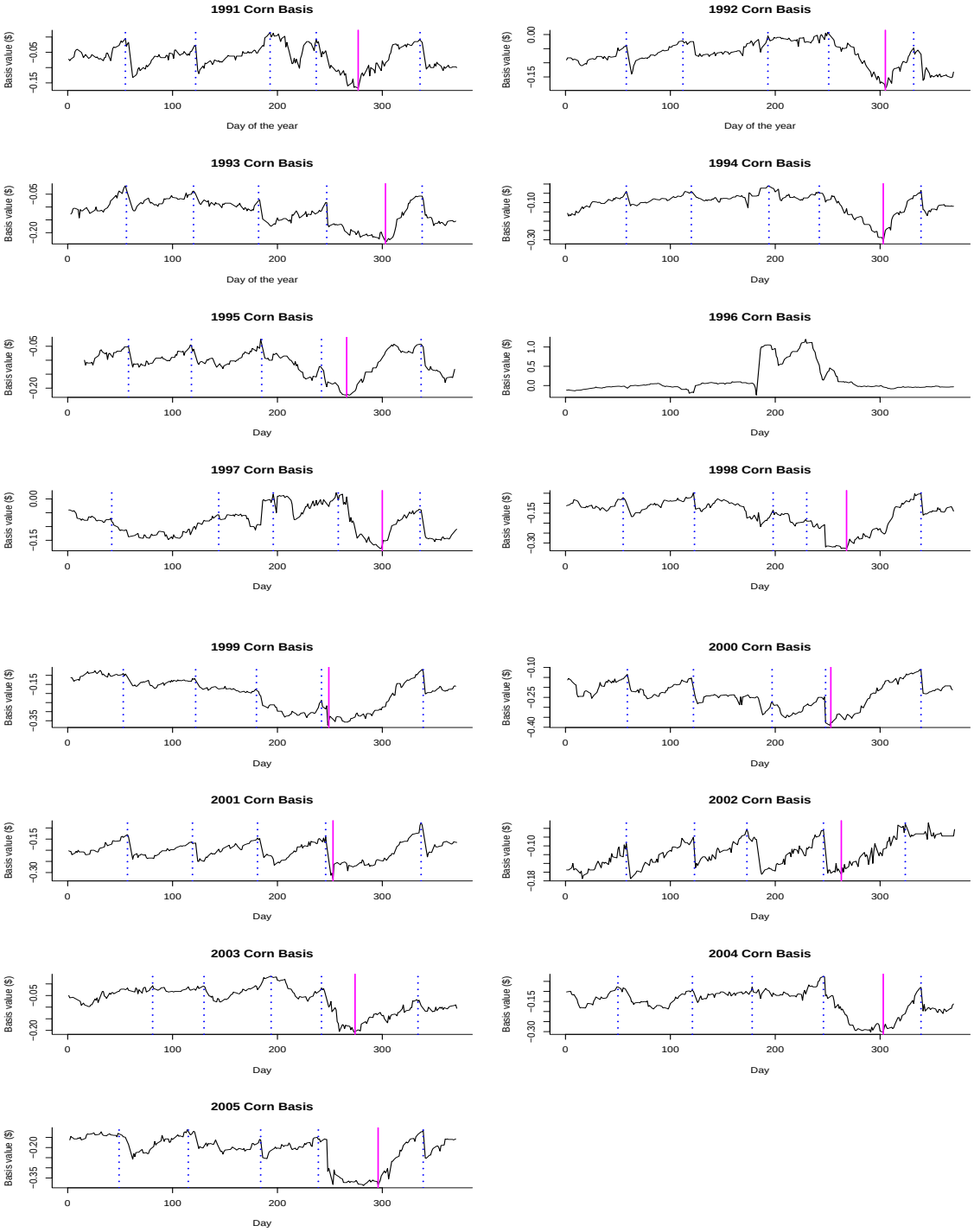

Figure 23: Corn basis plots from 1991 to 2005

Table 5: Dates and values of the corn basis at the extrema

	1st Maximum		2nd Maximum		3rd Maximum		4th Maximum		Minimum		5th Maximum	
Year	Date	Value ($)	Date	Value ($)	Date	Value ($)	Date	Value ($)	Date	Value ($)	Date	Value ($)
1991	2/25	-0.005	4/29	-0.0275	8/7	0.015	8/21	-0.01	9/30	-0.1675	11/27	-0.01
1992	2/28	-0.0375	4/20	-0.0225	7/8	-0.005	9/3	0.0075	10/27	-0.19	11/23	-0.0475
1993	2/25	-0.0175	4/28	-0.0375	6/29	-0.0725	8/31	-0.0825	10/26	-0.235	11/29	-0.0575
1994	2/28	-0.04	4/28	-0.04	7/8	-0.01	8/25	-0.0425	10/25	-0.31	11/30	-0.035
1995	2/28	-0.0525	4/25	-0.045	6/30	-0.025	8/25	-0.1225	9/22	-0.225	11/28	-0.0425
1997	2/11	-0.07	5/21	-0.0575	7/11	0.175	9/9	0.0225	10/21	-0.18	11/26	-0.0375
1998	2/25	-0.0725	4/30	-0.0475	7/13	-0.135	8/13	-0.1825	9/21	-0.3275	11/30	-0.0525
1999	2/23	-0.09	4/29	-0.1175	6/24	-0.185	8/26	-0.2375	9/1	-0.375	11/29	-0.0675
2000	2/29	-0.135	4/28	-0.14	7/12	-0.2725	8/31	-0.2475	9/7	-0.3775	11/29	-0.11
2001	2/27	-0.13	4/26	-0.1625	6/27	-0.145	8/30	-0.18	9/6	-0.3275	11/28	-0.0775
2002	2/28	-0.095	4/30	-0.08	6/19	-0.0625	8/30	-0.065	9/16	-0.165	11/15	-0.05
2003	3/20	-0.0075	5/7	-0.01	7/9	0.0325	8/26	-0.0175	9/26	-0.2075	11/24	-0.0675
2004	2/20	-0.0775	4/29	-0.095	6/30	-0.09	8/30	-0.025	11/3	-0.285	11/30	-0.08
2005	2/18	-0.13	4/22	-0.1125	6/29	-0.16	8/22	-0.15	10/18	-0.385	11/30	-0.115

August. The lowest basis occurs during September to October. Finally, the last local extremum occurs in the last week of November.

3.4.2.2 Soybean Basis

The soybean basis plots from 1991 to 2005 are shown in Figure 24. We distinguish a pattern in the soybean basis, but it is not as consistent as for the corn basis. In the soybean pattern, we identify five local maxima and one local minimum. They are indicated by the same lines used in identifying the corn basis's pattern. That is the first four local maxima are marked by the dashed lines on the left and the last local maxima is indicated by the right dashed line. The local minimum is marked by a straight line. The soybean basis for years 1997 and 2004 are regarded as the outliers in this study since they have the values much different from other years.

Table 6 provides dates and basis values at the local maxima and a local minimum of the soybean basis pattern. The first and second local extrema take place at the end of February and April, respectively. The third local maximum occurs during June

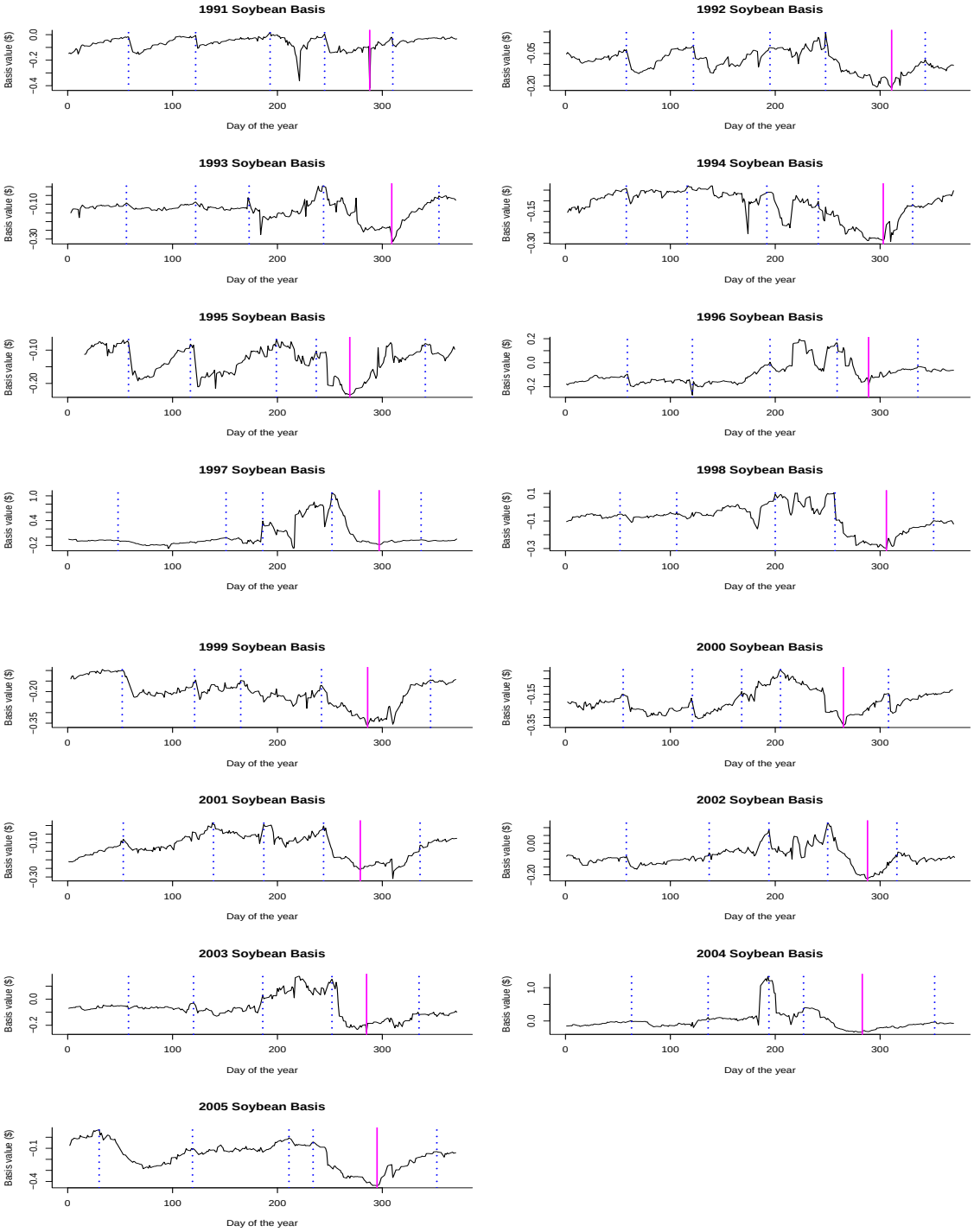

Figure 24: Soybean basis plots from 1991 to 2005

Table 6: Dates and values of the soybean basis at the extrema

	1st Maximum		2nd Maximum		3rd Maximum		4th Maximum		Minimum		5th Maximum	
Year	Date	Value ($)	Date	Value ($)	Date	Value ($)	Date	Value ($)	Date	Value ($)	Date	Value ($)
1991	2/28	-0.0175	4/30	-0.005	7/8	0.0175	8/29	0.005	10/10	-0.42	10/31	-0.02
1992	2/28	-0.0325	4/30	-0.0175	7/10	-0.0225	9/1	0.0475	11/2	-0.21	12/3	-0.0775
1993	2/26	-0.0925	4/30	-0.0875	6/18	-0.0625	8/27	0.005	11/1	-0.315	12/14	-0.0525
1994	2/28	-0.0425	4/25	-0.03	7/11	-0.06	8/30	-0.1875	10/25	-0.2875	11/22	-0.1175
1995	2/27	-0.0725	4/25	-0.0825	7/14	-0.0675	8/30	-0.105	9/22	-0.235	12/1	-0.08
1996	2/29	-0.095	4/26	-0.14	7/10	0.005	9/12	0.165	10/11	-0.1825	11/29	-0.0325
1997	2/11	-0.0675	5/28	-0.005	7/1	0.39	9/5	1.075	10/20	-0.18	11/28	-0.0625
1998	2/23	-0.05	4/14	-0.0375	7/15	0.095	9/9	0.105	10/28	-0.03	12/11	-0.0975
1999	2/24	-0.11	4/30	-0.145	6/11	-0.1475	8/26	-0.17	10/8	-0.36	12/6	-0.146
2000	2/24	-0.15	4/28	-0.175	6/14	-0.1325	7/20	0.01	9/18	-0.35	10/30	-0.145
2001	2/23	-0.08	5/15	0.015	7/9	0.0025	8/28	-0.0025	10/1	-0.26	12/4	-0.0975
2002	2/28	-0.0825	5/14	-0.06	7/9	0.085	9/3	0.13	10/9	-0.225	11/8	-0.055
2003	2/27	-0.05	4/28	-0.03	7/9	0.035	9/5	0.1525	10/7	-0.2525	11/25	-0.1
2004	3/2	0.01	5/13	0.105	7/8	1.34	8/12	0.4	10/5	-0.335	12/13	-0.0275
2005	1/31	0.0675	4/27	-0.1025	7/26	-0.005	8/18	-0.0425	10/17	-0.4375	12/12	-0.13

to July. The fourth local maximum consistently occurs in August or the first half of September. The local minimum mostly happens in October. Lastly, the fifth local extremum takes place during November to the first half of December. Moreover, we find that the patterns of corn and soybean basis are similar.

3.4.3 Functional Clustering Analysis

The functional model-based clustering technique outlined in Section 3.3.1 is applied to the commodity basis data. Similarly to yield forecasting, the commodity basis is forecasted yearly.

3.4.3.1 Corn Basis Forecast

We first perform the clustering analysis on the standardized corn basis data. As mentioned in Section 3.4.2.1, corn basis data in year 1996 is excluded from this analysis since it is an outlier for our 14-year period. For the method briefly discussed

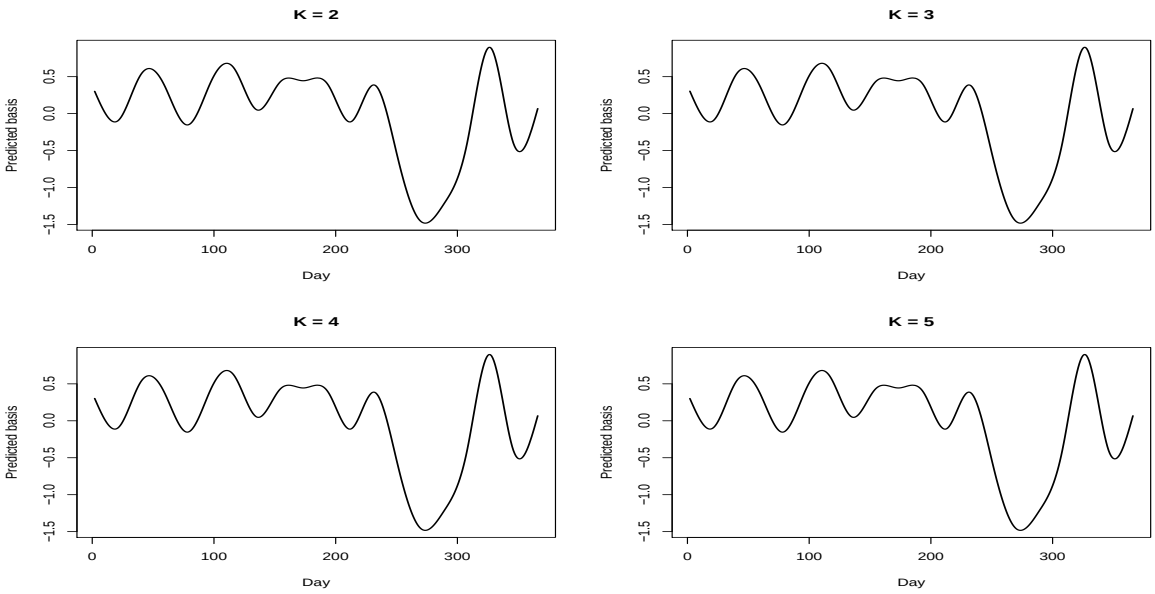

Figure 25: Comparison of 2005 corn basis forecasts when K = (2,3,4,5)

in Section 3.3.2 and developed by James and Sugar (2003), we need to specify the number of clusters, K, and the dimension of the spline basis, p. We determine a range of K from 2 to 5. Under the mixture likelihood framework (16), the predicted corn basis curves for each K are almost identical as depicted in Figure 25. This is because we have small number of periods to analyze. However, for a larger number of years, the number of clusters may play a significant role. By clustering the one-year corn basis curves, we hope to divide into years with unusual (high or low) production and years with normal production ($K = 2$) or into years with flood, years with drought, and years with normal weather condition ($K = 3$). There are techniques that can be used to determine the number of clusters, for example, Bayes factors (Kass and Raftery 1995), gap statistic (Tibshirani et al. 2001), and jump method (Sugar and James 2003) but here we choose a low value for K.

The dimension of the spline basis, p, plays an important role in the clustering analysis. Choosing low values of p results in low fitting quality. On the other hand, using high values of p results in overfitting. Figure 26 shows the prediction outcomes when p equals 10, 15, 20, and 25. In our further analysis, we choose $p = 20$ because we

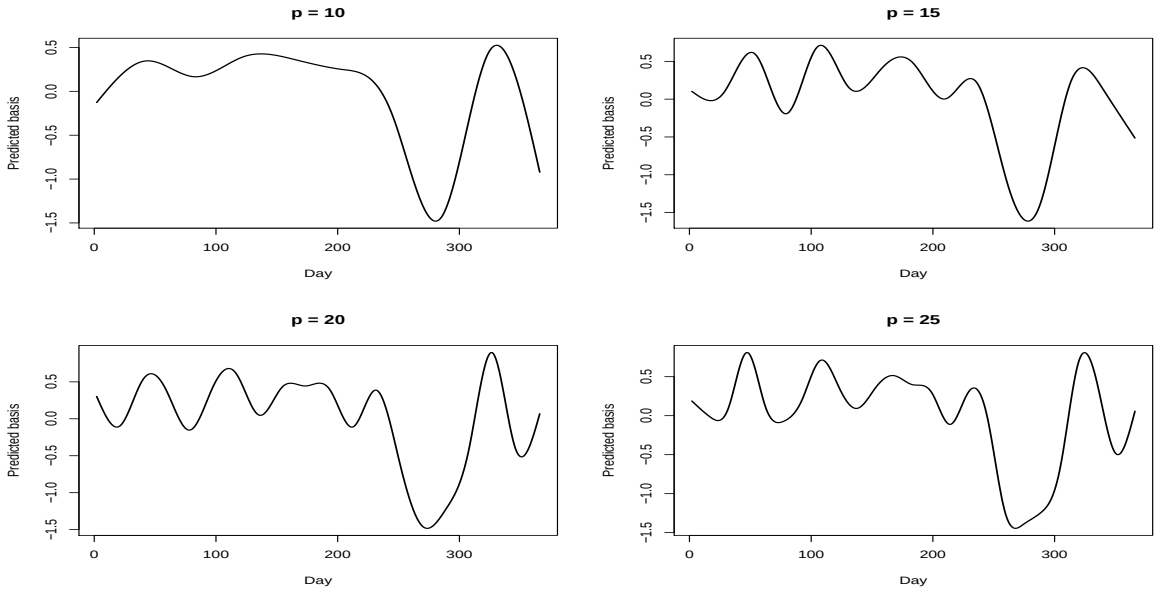

Figure 26: Comparison of 2005 corn basis forecasts when p = (10,15,20,25)

believe that at this smoothing level, we best balance the local and global variations.

3.4.3.2 Soybean Basis Forecast

We apply the functional clustering technique to the standardized soybean basis data as we did to the corn basis. The 1997 and 2004 soybean basis data are not considered in this study since they are outliers as discussed in Section 3.4.2.2. The ranges for K and p are investigated. Figure 27 shows the forecasted soybean basis curves when K equals 2, 3, 4, and 5. Again, these curves are almost the same so we choose a low value of K.

The tuning parameter that has a significant effect on the prediction is the dimension of the spline basis or the number of knots, p. As seen from Figure 28, $p = 20$ delivers a good fit that can capture the pattern without overfitting.

3.4.4 Prediction Confidence Band

The confidence band of the forecasted commodity basis, as defined in Section 3.3.2, is computed for $\alpha = 0.05$. The result for the corn basis forecast is illustrated in

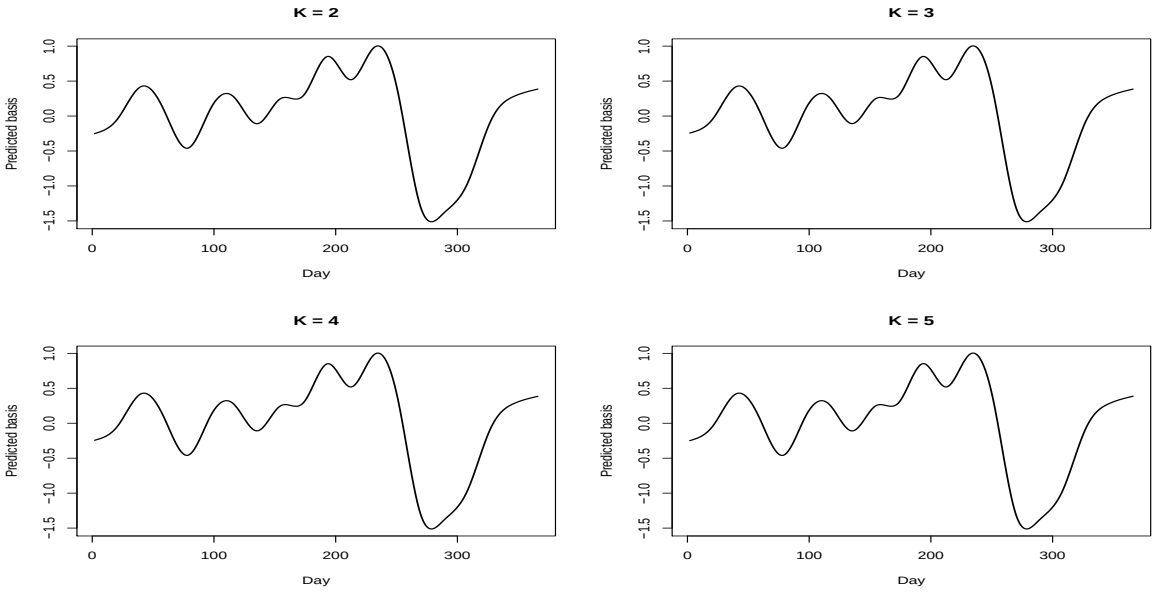

Figure 27: Comparison of 2005 soybean basis forecasts when K = (2,3,4,5)

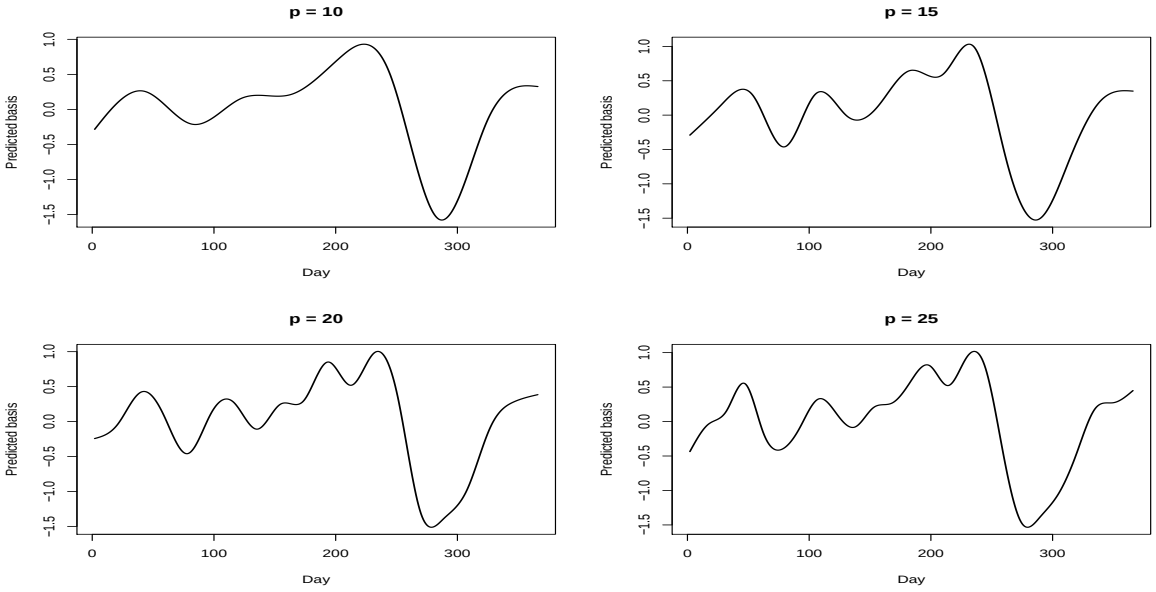

Figure 28: Comparison of 2005 soybean basis forecasts when p = (10,15,20,25)

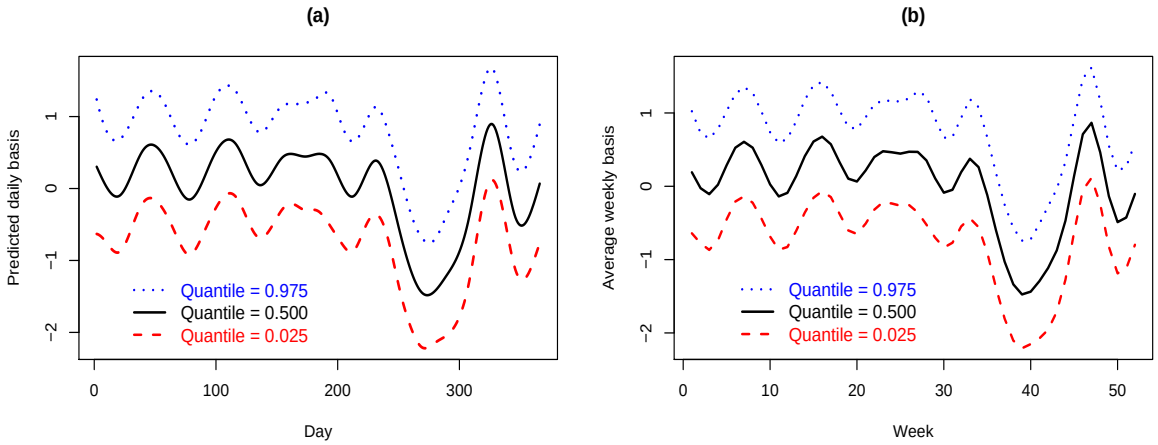

Figure 29: Plot of 95% pointwise prediction confidence bands of the predicted corn basis in daily (a) and average weekly (b)

Figure 29(a). Even though we have daily values available for corn basis historical data, it is more convenient to use weekly average values rather than daily values. This is because of the lack of synchronization between the dates across years. For example, January 2 was Friday in 2004 but Sunday in 2005. Moreover, the futures markets close on Saturday and Sunday so we do not have available data for weekend days. Consequently, the forecasted daily corn basis is averaged to a weekly value using forecasted data only from Monday to Friday. The confidence band in average weekly corn basis is shown in Figure 29(b). We can see that there is not much loss of information from aggregating daily corn basis values to weekly ones. The daily and average weekly pointwise confidence bands of the forecasted soybean basis are displayed in Figures 30(a) and 30(b), respectively.

3.4.5 Calibration

Since the commodity basis data is first standardized, the forecasted result will be on the standardized scale. The commodity basis on this scale provides information about the predicted pattern but not about the predicted values. In this section, we propose a simple calibration method.

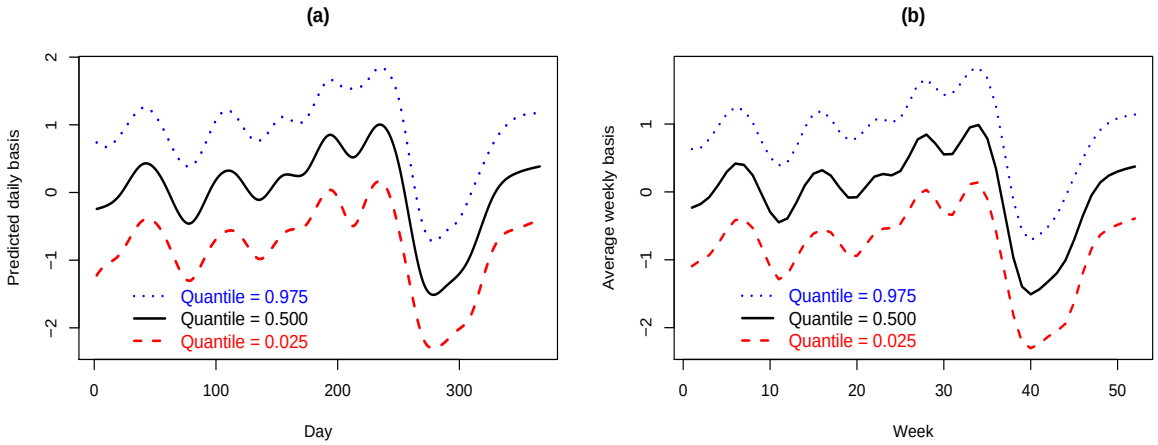

Figure 30: Plot of 95% pointwise prediction confidence bands of the predicted soybean basis in daily (a) and average weekly (b)

3.4.5.1 Corn Basis Calibration

The commodity basis patterns are on similar scales as those on their adjacent year. Therefore, we use the corn basis information from 2004 to calibrate the 2005 forecasted corn basis. We want to predict the corn basis starting at the time of planting, which is May for corn in Illinois. Therefore, we calibrate using the first four months of the forecasted year and the last eight months of the previous year to have a full year of reference corn basis data. In our calibration method, we first adjust for the mean difference between the current and previous years. We determine the difference between the means of the first four months of the observed corn basis curves in 2004 and 2005. Then we subtract the difference from the corn basis data in 2004 to shift the 2004 corn basis curve to the same level as one in 2005. Next, we re-scale the predicted pattern to the scale corresponding to the predicted year. The calibration result is illustrated in Figure 31. Most of the observed corn basis values are captured by the confidence band. However, there is still a large difference during weeks 35 to 45. This may come from the severe drought conditions in 2005, as mentioned in Section 2.4.5, which affected the corn production and hence the price.

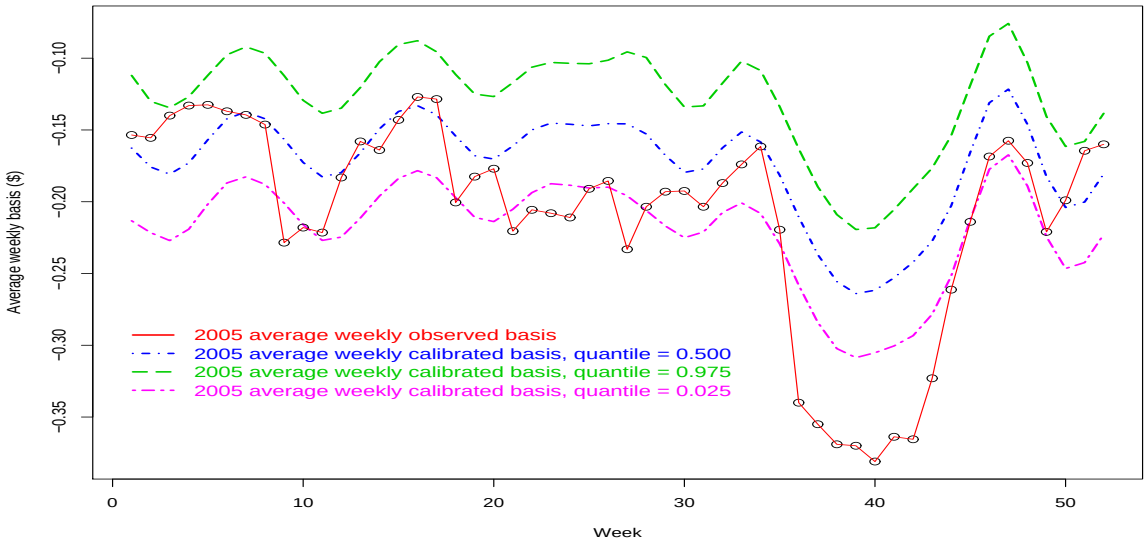

Figure 31: Plot of the 2005 average weekly observed corn basis and the calibrated prediction confidence band with difference adjusted

The proposed calibration technique works well in the years without severe condition. Figure 32 shows the calibrated band for corn basis in 2004. Most of the observed corn basis are contained within the band and there are only a few points that lie far from the band.

3.4.5.2 Soybean Basis Calibration

We calibrate the 2005 soybean basis forecast using the same calibration method described in Section 3.4.5.1. Since the soybean basis in 2004 is an outlier, as mentioned in Section 3.4.2.2, we use the soybean basis information in 2003 instead. Given that soybean in Illinois is first planted in May, we calibrate using the first four months of 2005 and the last eight months of 2003 as a reference soybean basis data. The calibration result is shown in Figure 33. Most of the observed soybean basis in 2005 are captured by the confidence band. There are several points that lie below the lower bound and a few points at the beginning that lie above the upper bound.

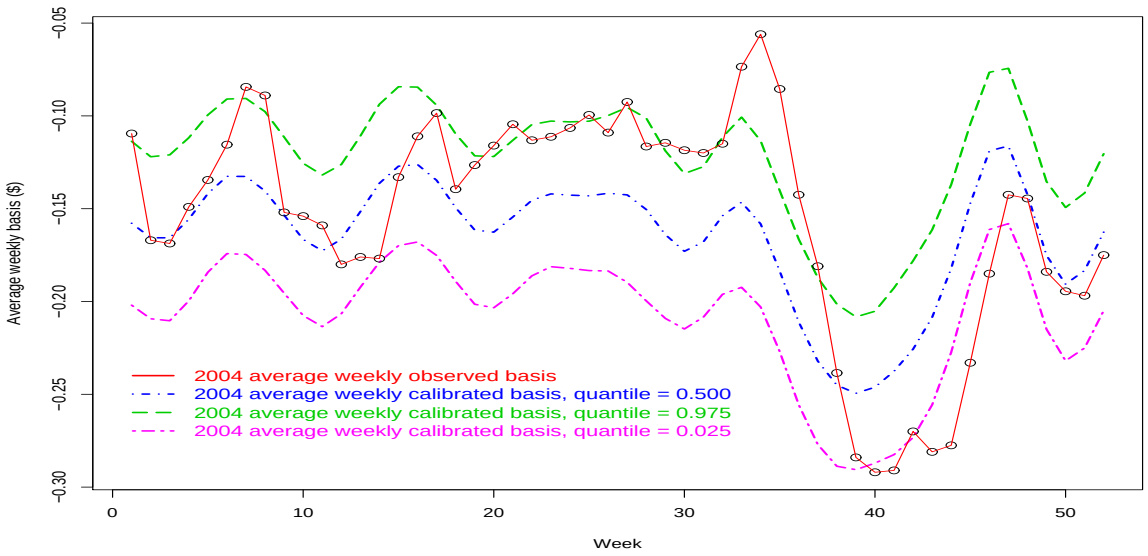

Figure 32: Plot of the 2004 average weekly observed corn basis and the calibrated prediction confidence band with difference adjusted

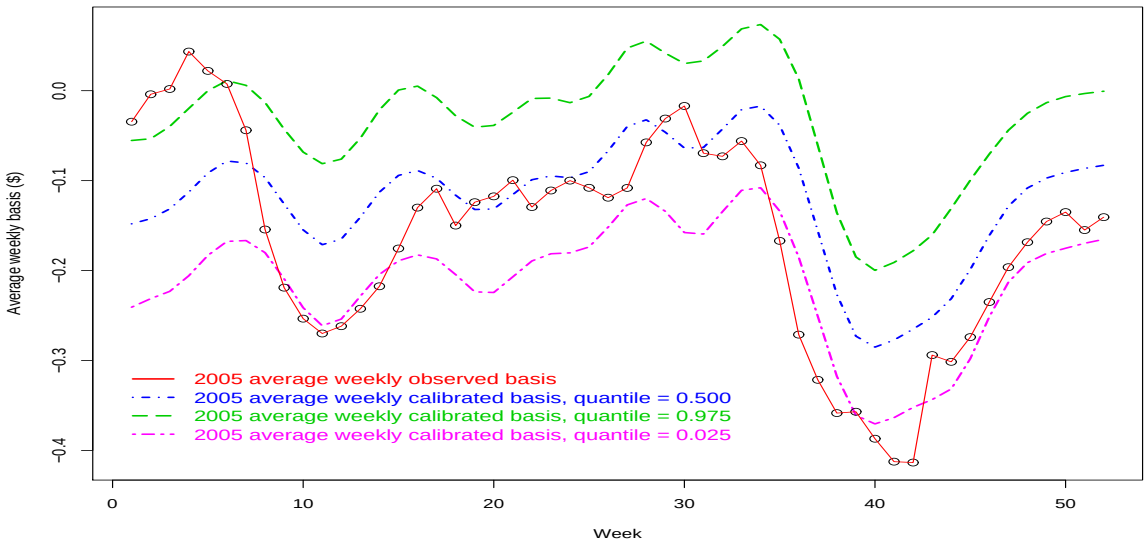

Figure 33: Plot of the 2005 average weekly observed soybean basis and the calibrated confidence band with difference adjusted

3.4.6 Forecasted Cash Price

In crop decision planning, we need to incorporate cash price but not the commodity basis. Cash price is calculated by adding the forecasted commodity basis to the expected futures price. For example, suppose we want to sell corn in the late September and the December futures contract (a nearby contract for September) is traded at \$2.265 per bushel. From the 2005 corn basis forecast, the corn basis in the last week of September (week 39) is -\$0.264 per bushel. Then the expected cash price is \$2.265 + (-\$0.264) = \$2.001 and the corresponding confidence interval is (\$1.9563, \$2.0457). Nevertheless, one has to bear in mind that the accuracy of the forecasted cash price depends not only on the commodity basis forecast but also on the futures price. The futures contract price changes continuously; even for the same delivery month the prices may not be the same if observed at different times. In the previous example, suppose that the day before our calculation, the December futures contract was traded at \$2.2575 per bushel. The expected cash price will change to \$2.2575 + (-\$0.264) = \$1.9935. The expected cash price for soybean can be calculated in a similar fashion.

3.5 Conclusions

This study explores the use of functional model-based analysis in the crop price forecasting. Under the futures-based model where expected cash price is equal to futures price plus expected commodity basis, we focus on commodity basis estimation. A multiple-year average technique is often used as a tool to compute the expected commodity basis. It is simple and provides relatively insightful results. However, it returns only the expectation of the commodity basis. Using only the expectation in decision making may lead to incorrect decisions. Our model formulation allows estimation of (pointwise) confidence band since it provides the density function of the commodity basis distribution.

We analyze and identify the pattern of the commodity basis information. Then, we apply the functional clustering analysis to the standardized commodity basis to estimate the basis distribution. We explore the appropriate values of the number of clusters and the dimension of the spline basis. Once the commodity basis distribution is estimated, we construct the confidence band and calibrate it to the normal scale. Lastly, the predicted commodity basis confidence interval is added to the futures contract price to obtain the forecasted cash price confidence interval.

From the numerical study results, both corn and soybean basis fluctuate with five local maxima and one local minimum. We also find that the number of clusters, K, under the mixture likelihood framework does not have much effect to the commodity basis estimation. The important factor is the dimension of the spline basis, p. Here, we decide to choose $p = 20$ for corn and soybean basis predictions since at this smoothing level, it delivers reasonable fits which balances the global and local variations. The proposed calibration method re-scale the standardized basis to its original scale. The calibrated pointwise confidence band can capture most of the observed commodity basis except for the corn basis in 2005, which has a large difference between the observed corn basis and the calibrated corn confidence band during the after-harvest periods. This discrepancy may come from the severe drought conditions in that year. In addition, we estimate the price with only a few years of data. If we were to have available data for a longer period of time, we might be able to cluster years with severe conditions like flood and drought. This would allow us to more accurately identify 2005 as a drought year and correctly account for this severe condition.

CHAPTER 4

CROP DECISION PLANNING

4.1 Introduction

A farmer has the objective of maximizing productivity. In order to achieve this goal, he has to carefully coordinate the decisions throughout the planning periods. However, this task is complicated since the decisions from one period affect the decisions in later periods. In addition, most agricultural decisions are irreversible. This means that once the farmer already makes decisions, he cannot change them. Furthermore, he has to consider many factors involving in the decision planning. One of them is resource limitation. The other generally come from the uncertainty in some key information such as weather conditions, yields, and prices.

We focus on the crop decision planning model and develop a stochastic linear programming model that incorporates constraints in resources such as land and labor as well as uncertainties in yields and prices. *The major contribution of our study is the development of a detailed planning model that determines the optimal decisions for crop selection, acreage allocation, planting and harvesting scheduling, storing, and selling.* In contrast to current approaches, the developed model utilizes the estimates from yield and price forecasting models. In addition, our proposed model examines the complete crop planning process while other studies focus on portions of the process. In addition to the proposed stochastic programming model, we also develop heuristic approaches based on greedy algorithms that provide feasible solutions at a low computational cost.

The remainder of this chapter is organized as follows. The next section reviews

relevant literature in crop decision planning. Section 4.3 details the problem definition. The stochastic programming model is presented in Section 4.4. The heuristic approaches are introduced in Section 4.5 and the numerical study of the proposed methods is illustrated and analyzed in Section 4.6. Finally, conclusions are given in Section 4.7.

4.2 Literature Review

Agricultural production planning is crucial in farm management. Carefully planned decisions can help increase profits. Butterworth (1985) discusses the Bedfordshire mixed cropping model which is used to determine decisions in the representative area under limited land, labor, and machinery. The main decisions are crop selection, beef cow retention, and spring-born calve fattening. He shows that the model can raise the gross margin by 28% compared to current practice.

Crop planning, a subset of agricultural production planning, primarily focuses on cropping activities. Glen (1987) provides a comprehensive literature survey in crop production models. Lowe and Preckel (2004) review the literature on crop planning, crop harvesting, and risk management based on operations research techniques. Frequently applied tools are linear programming, stochastic programming, risk programming, dynamic programming, and simulation.

Many crop planning models concentrate on crop selection and acreage allocation (Itoh et al. 2003, Sarker et al. 1997). One of the first studies in this area is discussed in Heady (1954). He demonstrates how to use a simple linear programming model to determine the acreage allocation among corn, potatoes, and oats. In his study, the primary objective is to maximize the profit subject to land, cash, and labor constraints. Apart from acreage allocation, many researchers account for uncertainty in their models (Glen 1987). Commonly used uncertain factors are price (Orazem and Miranowski 1986, Shonkwiler 1982, Shonkwiler and Emerson 1982) and yield

(Babcock 1990, Ethridge et al. 1975, Jones et al. 2002). Chavas and Holt (1990) and Marra and Carlson (1990) develop acreage response models for U.S. grain producers under expected utility maximization framework where the uncertainties are both price and yield.

Harvesting scheduling is a complex task since many factors have to be incorporated into the decision making. The main factors include harvesting capacity (machine and labor), weather conditions, and crop yields during the harvest season. Jiao et al. (2005) apply statistical and optimization techniques in harvesting scheduling decisions in Australian sugar production industry. The desired objective is to maximize the commercial cane sugar (CCS) and hence to maximize the profit under harvesting capacity limitation. A second-order polynomial regression model is used to fit CCS across a set of farms in the study. The estimated parameters are then utilized in a linear programming model to determine the proportion of tonnage of cane of each farm harvested in each harvest round. On average, profit increases by AUD 1.10 (or approximately USD 0.86) per ton of cane after adopting this method. Weather conditions are also widely studied. There are several researchers who have incorporated weather conditions in their models (Fokkens and Puylaert 1981, van Elderen 1980, Wilks et al. 1993).

Commodity storage is another important aspect, especially for storable crops, since products are harvested in a short period while demand is spread throughout the year. Finding an optimal storage policy has been a key objective for many researchers. The basic idea is simple. A producer should store grain and sell it after the harvest season when the prices are high enough to cover the storage costs and hopefully he will receive additional profit over selling at harvest. Fackler and Livingston (2002) propose a cutoff price function for risk neutral farmers. The optimal decision rule is based on a cutoff price; selling all stocks if the market price exceeds the cutoff price, otherwise keeping the stocks. The problem scope is only at the on-farm produced

crops, which makes sales decision an irreversible action. Central Illinois soybean average bid prices are used to evaluate the performance of this policy and this strategy results in an additional return of 35-55 cents per bushel over cash sale at harvest. Lai et al. (2003) extend the model of Fackler and Livingston (2002) to the risk-averse analysis using stochastic dynamic programming in discrete-time framework. They find that the risk-averse farmers will spread the sales over the storage season rather than selling everything or nothing in the case of risk-neutral producers. Many studies are carried out to determine the impact on storage decisions caused by taxes (McNew and Gardner 1999, Tronstad and Taylor 1991), U.S. farm policy (Lence and Hayes 2002), and futures market (Netz 1995, Sexauer 1977).

This research differs from the above models in that the whole crop planning process is determined, including crop selection, acreage allocation, planting and harvesting scheduling, and storage and selling decisions. Yield and price are regarded as uncertain factors in the model. Hence, this research problem can be classified as a *sequential decision problem under uncertainty*. One of the common approaches for this type of problem is stochastic programming.

Stochastic programming in crop decision planning has received wide attention in agricultural literature. Cocks (1968) studies a profit maximization problem where the farmer has to allocate land between wheat and sugar beets. Labor and gross margins are uncertain and become known after making the allocation decision. He shows that stochastic programming provides a higher expected profit than linear programming. Maatman et al. (2002) formulate farmers' sequential decision making under rainfall uncertainty in Burkina Faso as a two-stage stochastic model with recourse. The first stage decisions are made after observing the first rains. Nevertheless, there is still an uncertainty in rainfall during the growing season. The second stage decisions are made when the latter rainfall is observed. The objective function is designed such that it gives the highest priority to minimize deficits of nutrients during the planning

period and the lowest priority to the revenues. A more recent application of stochastic programming can be found in Kazaz (2004). He determines the two-stage stochastic program with recourse in the olive oil industry. The random variables are yield and demand. The first stage decision involves the amount of farm space to be leased while the second stage decision is the amount of olive oil produced from internally grown and purchased olives.

Overall, *our research focuses on farm-level crop planning model under a stochastic programming framework by considering the sequence of decisions made by farmers. The objective is to maximize the expected profit under yield and price uncertainties.* This model accounts for constraints in land, labor, and crops' minimum requirement amount as well as the limitation on the planting and harvesting periods of each crop. In addition, heuristic approaches are developed to solve the problem when the problem size is very large.

4.3 Problem Definition

Consider a farmer who plans to grow storable crops in the coming year. We assume that he already owns the land and the necessary machines and equipment used in cultivation. Therefore, investment in technology is not considered. The farmer has to make many decisions during the cropping periods in order to achieve his desired goal - to maximize expected total profit. The revenues come from selling his products while the costs associated with his operations include:

1. Planting costs
2. Harvesting costs
3. Storage costs
4. Transportation costs.

These costs are assumed to be fixed and known in advance. The cost information can be gathered from previous cropping years or from published agricultural reports. The crop decision planning model is then designed to help the farmer answer the following questions:

1. Which crops should be grown this year?
2. How much acreage should he allocate to each selected crop?
3. When should he plant the crops?
4. How much land should he plant for each crop in each planting period?
5. When should he harvest the crops?
6. How much land should he harvest for each crop in each harvesting period?
7. Should he sell the crops at harvest or keep them for later sale?
8. What are the best times to sell his crops?

In this research, we make the following assumptions. First, we do not consider investment since we focus on short term decision planning. Therefore, we rule out the option to buy or rent more land. Second, we assume that no additional farm workers are hired or fired over the planning periods and the farm is solely operated by the owner. This is because most of the farms in the representative area, Illinois, have a single operator. Third, since each crop has its own planting and harvesting periods, we assume that growing crops before or after their planting periods will severely reduce the yields. Moreover, heavy losses will incur if harvest operations are not performed during their harvesting periods. However, small losses may occur if crops are harvested after the middle of the harvesting periods. In order to achieve effective production, the farmer must carefully allocate the labor in each planting

and harvesting period of each crop. Finally, we assume that there are uncertainties in yields and prices. These uncertainties will be forecasted and integrated into the planning model.

In our study, the planning horizon covers the duration from the beginning of the planting season through the end of year, which can be divided into five decision stages as illustrated in Figure 34. Stage 1 corresponds to the planting periods. Stages 2 and 3 correspond to the first and second halves of the harvesting periods, respectively. The time periods after harvesting periods are considered as stage 4 except for the end of year which is regarded as stage 5. We assume that the farmer can sell his products only at the middle of the harvest season, the end of the harvest season, and the end of year and no crop is carried over to the next year. The reasons for choosing these selling periods are as follows. First, after the middle of the harvest season, crop yield may decrease but crop price may increase. Thus, we would like to investigate the decisions that trade off between two alternatives. The first one is to harvest early and sell at the middle of the harvest season, which usually has low selling price. The second alternative is to harvest and sell late, which has a chance to have yield loss but high selling price. Second, we regard the end of the harvest season as representative of the selling periods during the harvest season. Finally, the end of year is the end of the planning horizon, which is the last period that the farmer can sell his products. In addition, there is a crop's minimum requirement amount to be satisfied at the end of the harvest season which will be used on the farm. This requirement amount can be supplied from the farm production and/or purchased from the market. The purchasing price is usually higher than the selling price due to the seller's profit margin. Since this research focuses on farm planning, we assume that the farmer can purchase crops just to satisfy the minimum requirement amount and is not allowed to speculate on these crops.

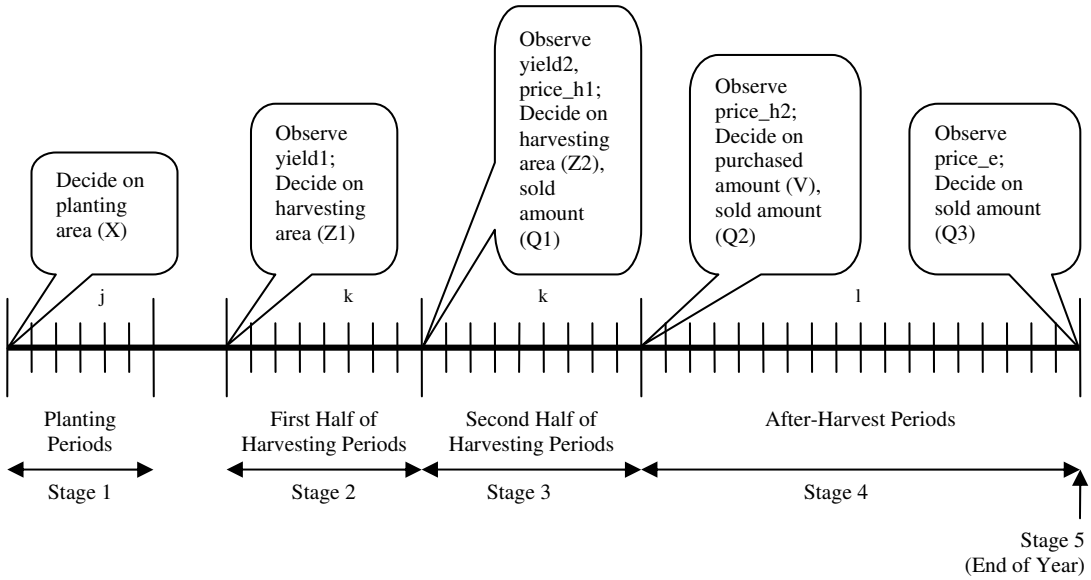

Figure 34: Time line for planning horizon of crop decision planning

There are five uncertainties considered in our model. The first one is the yield at the beginning of the harvest season (*yield1*). The second and the third uncertainties correspond to the yield (*yield2*) and price (*price_h1*) at the middle of the harvest season, respectively. The fourth one is the price at the end of the harvest season (*price_h2*). The last uncertainty is the price at the end of year (*price_e*).

4.4 Stochastic Programming Model

In this section, we discuss the decision variables associated with each stage within the planning horizon. These decisions are made at the beginning of the corresponding stages. Then we present the structure of stochastic programming model and discuss the components of the objective function and constraints.

4.4.1 Decision Variables

Let index i correspond to crops. Let J, K, and L be the sets of planting periods, harvesting periods, and after-harvest periods, respectively. Each of these periods has

several micro-periods which are denoted by j for micro-periods in J, k for micro-periods in K, and l for micro-periods in L. Let s_1, s_2, s_3, s_4, and s_5 be the observed outcomes of yield1, yield2, price_h1, price_h2, and price_e, respectively. For example, s_1 may be low yield1, medium yield1, or high yield1 when yield1 has three outcomes - low, medium, and high. We define the decision variables in each stage as follows.

Stage 1: The farmer makes decisions regarding the crop selection and acreage allocation as well as the planting schedule. *At this stage, the yields and prices of crops are still unknown.* The first-stage decision variable is defined as

$X(i, j)$: Planting area (in acre) for crop i in period j.

Stage 2: *'Yield1' is observed at the beginning of the harvest season.* It is assumed to be constant throughout this stage. The grower decides the harvesting schedule and also takes into account that the yield may be decreased in the second half of the harvest season. The second-stage decision variable is defined as

$Z1(i, k|s_1)$: Harvesting area (in acre) for crop i in period k given outcome s_1.

Stage 3: The first selling decision is made at the middle of the harvest season. *At this moment, 'yield2' and 'price_h1' are known.* The current yield may be the same or less than the yield in stage 2 but it is assumed to be constant throughout this stage. The producer decides to sell the crops that are already harvested at the current price or keep them in storage for later sale. In addition, he plans the harvesting schedule for crops that are not collected during the previous stage. Therefore, the third-stage decision variables are defined as

$Z2(i, k|s_1, s_2, s_3)$: Harvesting area (in acre) for crop i in period k given outcomes s_1, s_2, s_3;

$Q1(i|s_1, s_2, s_3)$: Amount of crop i (in bushel) sold at the middle of the harvest season given outcomes s_1, s_2, s_3.

Stage 4: *'Price_h2' is observed at the end of harvest season.* The farmer decides

how to satisfy the minimum requirement amount and whether or not to sell the products at this price or to sell them at the end of year. The decision variables in the fourth-stage are defined as

$V(i\|s_1, s_2, s_3, s_4)$: Amount of crop i (in bushel) bought at the end of the harvest season given outcomes s_1, s_2, s_3, s_4;
$Q2(i\|s_1, s_2, s_3, s_4)$: Amount of crop i (in bushel) sold at the end of the harvest season given outcomes s_1, s_2, s_3, s_4.

Stage 5: Any crops left in the storage are sold at the end of year at the observed 'price_e'. The fifth-stage decision variable is then defined as

$Q3(i\|s_1, s_2, s_3, s_4, s_5)$: Amount of crop i (in bushel) sold at the end of year given outcomes s_1, s_2, s_3, s_4, s_5.

4.4.2 Model

A stochastic programming model leads in a policy or strategy. This policy specifies the set of decisions that the farmer should take in every stage for any possible scenario that will arise in the future. In addition, it provides the expected profit from using this policy in the long run.

The sets, parameters, observations, prior information, variables, and mathematical formulation are defined as follows.

Sets

I = Set of crops,

J = Set of all planting periods,

J_i = Set of planting periods for crop i,

K_1 = Set of periods in the first half of the harvest season,

K_2 = Set of periods in the second half of the harvest season,

K_{1i} = Set of periods in the first half of the harvest season for crop i,

K_{2i} = Set of periods in the second half of the harvest season for crop i,

L_i = Set of after-harvest periods for crop i,

$\mathbb{SP}1_i$ = First selling time period for crop i, defined as the middle of the harvest season,

$\mathbb{SP}2_i$ = Second selling time period for crop i, defined as the end of the harvest season,

$\mathbb{SP}3_i$ = Third selling period for crop i, defined as the end of year,

S_1 = Set of yield outcomes at the beginning of the harvest season (yield1),

S_2 = Set of yield outcomes at the middle of the harvest season (yield2),

S_3 = Set of price outcomes at the middle of the harvest season (price_h1),

S_4 = Set of price outcomes at the end of the harvest season (price_h2),

S_5 = Set of price outcomes at the end of year (price_e).

Parameters

Lnd = Available land (acre),

r = Interest rate per period (%),

λ = Penalty for producing less than minimum requirement, expressed as percentage of selling price at the end of the harvest season,

$D(i)$ = Minimum requirement for crop i (bushel),

$WPL(j)$ = Available labor in planting period j (hour),

$WHA(k)$ = Available labor in harvesting period k (hour),

$wp(i)$ = Labor needed (in hour/acre) to plant an acre of crop i,

$wh1(i|s_1)$ = Labor needed (in hour/acre) to harvest an acre of crop i during the first half of the harvest season given outcome s_1,

$wh2(i|s_1, s_2)$ = Labor needed (in hour/acre) to harvest an acre of crop i during the second half of the harvest season given outcomes s_1, s_2,

$c(i, j)$ = Planting cost of crop i in period j (\$/acre) ,

$h(i, k)$ = Harvesting cost per acre of crop i in period k (\$/acre),

$sh(i, k)$ = Storage cost during the harvest season per unit of crop i in period k (\$/bushel),

$sa(i, l)$ = Storage cost after the harvest season per unit of crop i in period l (\$/bushel),

$t(i)$ = Transportation cost per unit of crop i (\$/bushel),

$FP(j)$ = Compounding factor for planting period j,

$FH(k)$ = Compounding factor for harvesting period k,

$FS(l)$ = Compounding factor for after-harvest period l,

$FQ_1(i)$ = Compounding factor for amount of crop i sold at the middle of the harvest season,

$FQ_2(i)$ = Compounding factor for amount of crop i sold at the end of the harvest season,

$FQ_3(i)$ = Compounding factor for amount of crop i sold at the end of year.

Observations

$yield1(i|s_1)$ = Yield of crop i (in bushel/acre) at the beginning of the harvest season given outcome s_1,

$yield2(i|s_1, s_2)$ = Yield of crop i (in bushel/acre) at the middle of the harvest season given outcomes s_1, s_2,

$price\_h1(i|s_3)$ = Price of crop i (in \$/bushel) at the middle of the harvest season given outcome s_3,

$price\_h2(i|s_4)$ = Price of crop i (in \$/bushel) at the end of the harvest season given outcome s_4,

$price\_e(i|s_5)$ = Price of crop i (in \$/bushel) at the end of year given outcome s_5.

Prior Information

$p_1(s_1)$ = Probability that outcome s_1 will occur,

$p_2(s_2, s_3|s_1)$ = Conditional probability that outcomes s_2 and s_3 will occur given outcome s_1,

$p_3(s_4|s_1, s_2, s_3)$ = Conditional probability that outcome s_4 will occur given outcomes s_1, s_2, s_3,

$p_4(s_5|s_1, s_2, s_3, s_4)$ = Conditional probability that outcome s_5 will occur given outcomes s_1, s_2, s_3, s_4.

Variables

$X(i, j)$ = Planting area (in acre) for crop i in period j,

$Z1(i, k|s_1)$ = Harvesting area (in acre) for crop i in period k during the first half of the harvest season given outcome s_1,

$Z2(i, k|s_1, s_2, s_3)$ = Harvesting area (in acre) for crop i in period k during the second half of the harvest season given outcome s_1, s_2, s_3,

$UH1(i, k|s_1)$ = Amount of crop i (in bushel) stored in period k during the first half of the harvest season, given outcome s_1,

$UH2(i, k|s_1, s_2, s_3)$ = Amount of crop i (in bushel) stored in period k during the second half of the harvest season, given outcomes s_1, s_2, s_3,

$UA(i, l|s_1, s_2, s_3, s_4)$ = Amount of crop i (in bushel) stored in period l given outcomes s_1, s_2, s_3, s_4,

$V(i|s_1, s_2, s_3, s_4)$ = Amount of crop i (in bushel) bought at the end of the harvest season given outcomes s_1, s_2, s_3, s_4,

$Q1(i|s_1, s_2, s_3)$ = Amount of crop i (in bushel) sold at the middle of the harvest season given outcomes s_1, s_2, s_3,

$Q2(i|s_1, s_2, s_3, s_4)$ = Amount of crop i (in bushel) sold at the end of the harvest season given outcomes s_1, s_2, s_3, s_4,

$Q3(i|s_1, s_2, s_3, s_4, s_5)$ = Amount of crop i (in bushel) sold at the end of year given outcomes s_1, s_2, s_3, s_4, s_5.

The stochastic linear programming model for crop decision planning can be formulated as:

Objective function:

Maximize

$$\left.\begin{array}{l} \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\left[\sum_{i\in I} FQ_1(i)\cdot price\_h1(i|S_3)\cdot Q1(i|S_1,S_2,S_3)\right] \\ +\ \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\mathbb{E}_{S_4}\left[\sum_{i\in I} FQ_2(i)\cdot price\_h2(i|S_4)\cdot Q2(i|S_1,S_2,S_3,S_4)\right] \\ +\ \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\mathbb{E}_{S_4}\mathbb{E}_{S_5}\left[\sum_{i\in I} FQ_3(i)\cdot price\_e(i|S_5)\cdot Q3(i|S_1,S_2,S_3,S_4,S_5)\right] \end{array}\right\} \text{Revenues}$$

$$\left.-\ \sum_{i\in I}\sum_{i\in J} FP(j)\cdot c(i,j)\cdot X(i,j)\right\} \text{Planting costs}$$

$$\left.\begin{array}{l} -\ \mathbb{E}_{S_1}\left[\sum_{i\in I}\sum_{k\in K_1} FH(k)\cdot h(i,k)\cdot Z1(i,k|S_1)\right] \\ -\ \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\left[\sum_{i\in I}\sum_{k\in K_2} FH(k)\cdot h(i,k)\cdot Z2(i,k|S_1,S_2,S_3)\right] \end{array}\right\} \text{Harvesting costs}$$

$$\left.-\ \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\mathbb{E}_{S_4}\left[\sum_{i\in I} FQ_2(i)\cdot(1+\lambda)\cdot price\_h2(i|S_4)\cdot V(i|S_1,S_2,S_3,S_4)\right]\right\} \text{Penalty costs} \quad (20)$$

$$
\left.
\begin{aligned}
&- \mathbb{E}_{S_1}\left[\sum_{i\in I}\sum_{k\in K_1} FH(k)\cdot sh(i,k)\cdot UH1(i,k|S_1)\right] \\
&- \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\left[\sum_{i\in I}\sum_{k\in K_2} FH(k)\cdot sh(i,k)\cdot UH2(i,k|S_1,S_2,S_3)\right] \\
&- \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\mathbb{E}_{S_4}\left[\sum_{i\in I}\sum_{i\in L} FS(l)\cdot sa(i,l)\cdot UA(i,l|S_1,S_2,S_3,S_4)\right]
\end{aligned}
\right\} \text{Storage costs}
$$

$$
\left.
\begin{aligned}
&- \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\left[\sum_{i\in I} FQ_1(i)\cdot t(i)\cdot Q1(i|S_1,S_2,S_3)\right] \\
&- \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\mathbb{E}_{S_4}\left[\sum_{i\in I} FQ_2(i)\cdot t(i)\cdot Q2(i|S_1,S_2,S_3,S_4)\right] \\
&- \mathbb{E}_{S_1}\mathbb{E}_{S_2}\mathbb{E}_{S_3}\mathbb{E}_{S_4}\mathbb{E}_{S_5}\left[\sum_{i\in I} FQ_3(i)\cdot t(i)\cdot Q3(i|S_1,S_2,S_3,S_4,S_5)\right]
\end{aligned}
\right\} \text{Transportation costs}
$$

Constraints:

For all $s_1 \in S_1, s_2 \in S_2, s_3 \in S_3, s_4 \in S_4, s_5 \in S_5$:

Land Constraint:

$$\sum_{i\in I}\sum_{j\in J} X(i,j) = Lnd \tag{21}$$

Labor Constraints:

$$\sum_{i\in I} wp(i)\cdot X(i,j) \le WPL(j), \quad j \in J \tag{22}$$

$$\sum_{i\in I} wh1(i|s_1)\cdot Z1(i,k|s_1) \le WHA(k), \quad k \in K_1 \tag{23}$$

$$\sum_{i\in I} wh2(i|s_1,s_2)\cdot Z2(i,k|s_1,s_2,s_3) \le WHA(k), \quad k \in K_2 \tag{24}$$

Harvesting Area Constraint:

$$\sum_{k\in K_1} Z1(i,k|s_1) + \sum_{k\in K_2} Z2(i,k|s_1,s_2,s_3) \le \sum_{j\in J} X(i,j), \quad i \in I \tag{25}$$

Inventory Balance Constraints:

$$UH1(i,k|s_1) = UH1(i,k-1|s_1) + yield1(i|s_1)\cdot Z1(i,k|s_1), \quad i \in I,$$
$$k \in K_{1i} \tag{26}$$

$$UH2(i,k|s_1,s_2,s_3) = UH1(i,k-1|s_1) + yield2(i|s_1,s_2)\cdot Z2(i,k|s_1,s_2,s_3)$$
$$- Q1(i|s_1,s_2,s_3), \quad i \in I, k = \mathbb{SP}1_i \tag{27}$$

$$UH2(i,k|s_1,s_2,s_3) = UH2(i,k-1|s_1,s_2,s_3)$$
$$+ yield2(i|s_1,s_2)\cdot Z2(i,k|s_1,s_2,s_3), \quad i \in I,$$
$$k \in K_{2i} \setminus \{\mathbb{SP}1_i, \mathbb{SP}2_i\} \tag{28}$$

$$D(i) + UA(i,l|s_1,s_2,s_3,s_4) = UH2(i,k-1|s_1,s_2,s_3)$$
$$+ yield2(i|s_1,s_2)\cdot Z2(i,k|s_1,s_2,s_3) - Q2(i|s_1,s_2,s_3,s_4)$$
$$+ V(i|s_1,s_2,s_3,s_4) \quad i \in I, k = l = \mathbb{SP}2_i \tag{29}$$

$$UA(i,l|s_1,s_2,s_3,s_4) = UA(i,l-1|s_1,s_2,s_3,s_4),\ i \in I, l \in L_i \setminus \{\mathbb{SP}2_i, \mathbb{SP}3_i\} \tag{30}$$

$$Q3(i|s_1,s_2,s_3,s_4,s_5) = UA(i,l-1|s_1,s_2,s_3,s_4), \quad i \in I, l = \mathbb{SP}3_i \tag{31}$$

No-speculation Constraint:

$$V(i|s_1,s_2,s_3,s_4) \leq D(i), \quad i \in I \tag{32}$$

Restriction Constraints:

$$\sum_{j \in J \setminus J_i} X(i,j) = 0, \quad i \in I \tag{33}$$

$$\sum_{k \in K_1 \setminus K_{1i}} Z1(i,k|s_1) = 0, \quad i \in I \tag{34}$$

$$\sum_{k \in K_2 \setminus K_{2i}} Z2(i,k|s_1,s_2,s_3) = 0, \quad i \in I \tag{35}$$

Nonnegativity Constraints:

$$X(i,j) \geq 0, i \in I, j \in J \tag{36}$$

$$Z1(i,k|s_1),\ UH1(i,k|s_1) \geq 0, i \in I, k \in K_{1i} \tag{37}$$

$$Z2(i,k|s_1,s_2,s_3),\ UH2(i,k|s_1,s_2,s_3) \geq 0, i \in I, k \in K_{2i} \tag{38}$$

$$UA(i,l|s_1,s_2,s_3,s_4) \geq 0, i \in I, l \in L_i \tag{39}$$

$$V(i|s_1,s_2,s_3,s_4),\ Q1(i|s_1,s_2,s_3),\ Q2(i|s_1,s_2,s_3,s_4),$$
$$Q3(i|s_1,s_2,s_3,s_4,s_5) \geq 0, i \in I \tag{40}$$

The objective function of the model in equation (20) maximizes expected future value of the overall profit at the end of planning horizon. It is computed by subtracting expected future values of costs from expected future values of revenues. The expected revenues are generated from the crop sold. The costs associated in this model are planting costs, harvesting costs, penalty costs, storage costs, and transportation costs. The overall profit can be viewed as a summation of the profit obtained from each crop.

The constraints are given as follows. Constraint (21) restricts the overall planting area to be less than or equal to the available land. Constraints (22)-(24) limit labor hours used in planting and harvesting periods not to exceed given labor hours in those periods. Constraint (25) bounds the total harvesting area of crop i by the total planting area of that crop. Constraints (26), (28), and (30) maintain the inventory balance in the periods without selling, while constraints (27), (29), and (31) determine the inventory balance in the selling periods. Note that, constraint (29) implies that the minimum requirement can be supplied by the crop available on hand and/or crop purchased from the market. Constraint (32) rules out the speculation opportunity by setting the upper bound on the amount of crops a farmer could purchase from the market. Constraints (33)-(35) state that each crop can be planted and harvested only in its planting and harvesting periods. Finally, constraints (36)-(40) force all variables to be nonnegative.

In our research, we consider a linear version of stochastic programming called stochastic linear programming (**SLP**). From the above model, it can be seen that

the SLP model is similar to linear deterministic optimization model except that the former model has some unknown parameters where the latter is formulated only with known parameters. If the probability distributions of the unknown parameters can be estimated then the SLP model can be solved in the same way as linear deterministic optimization by maximizing the expectation given in the objective function. Hence, by employing prior probabilities of the unknown parameters or uncertainties, we can formulate our problem as a linear programming (LP), which can be solved by the simplex method.

4.5 Heuristic Approaches

Even though stochastic programming is a commonly used technique to solve the sequential decision making under uncertainty, it has a significant drawback. The size of the computational effort increases exponentially with the number of decision stages (Yu et al. 2003, Topaloglou 2004). The crop decision planning with a moderate number of stages and outcomes at each stage can become a very large-scale optimization problem. For example, the problem with ten stages where each stage has ten possible outcomes will have ten thousand million possible scenarios. We therefore develop simple heuristic models which give feasible solutions in a reasonable amount of time.

The heuristic models we propose here are based on greedy algorithms which make the optimal decisions based only on information available on hand and do not consider the effect that these decisions may cause in the future. In these heuristics, we apply the greedy algorithms to the deterministic version of the SLP model ((20)-(40)), which does not have the expectation operators in the objective function (20). The linear deterministic model is solved for each stage using the information available up to that stage.

The difference between our heuristic approaches and the stochastic programming approach is that the latter yields the policy or strategy for every possible scenario

while these heuristics only provide the solution for a particular scenario. This may seem to be a drawback to the heuristic approaches because they do not provide a complete policy whereas the stochastic programming approach does. However, in practice, it may not be feasible to compute the strategy for every scenario in advance, especially when the number of scenarios is large. In addition, the decision makers have to observe the uncertainties before taking actions, which are provided by the policy, in the next stage. Moreover, when all uncertainties are observed, what decision makers use is the set of decisions for the scenario that actually occurs. There will be other scenarios that are not considered. The heuristic models offer a faster way to compute the solution for occurred scenario and bypass the calculation for scenarios that do not occur. Hence, our heuristics are not much different from the stochastic programming approach when used in practice. Since our heuristics are solved along the way during the planning periods, we call them as *solve-on-the-fly* heuristics.

The reason that we choose the greedy approach is that it is easy to implement. In addition, since it is formulated as an ordinary LP which maximizes the linear objective function, it can be solved much faster than SLP. However, it has a limitation - it can provide only the feasible solutions that sometimes can be much worse than the optimal solution generated by SLP.

The general structure of the greedy algorithms is illustrated in Figure 35. The *solution set* is used to keep the model solution. The *realization set* is used to keep the observed outcomes of uncertainties. The *pre-specified criterion* is the rule used to determine the outcomes of the future uncertainties. For example, we may choose the most likely outcome as a representative of the uncertainty. The detailed steps of the algorithms are given in Appentix A.

We propose three heuristics which share the same algorithm structure but have different pre-specified criteria to determine the outcomes of the future uncertainties. The proposed heuristics are:

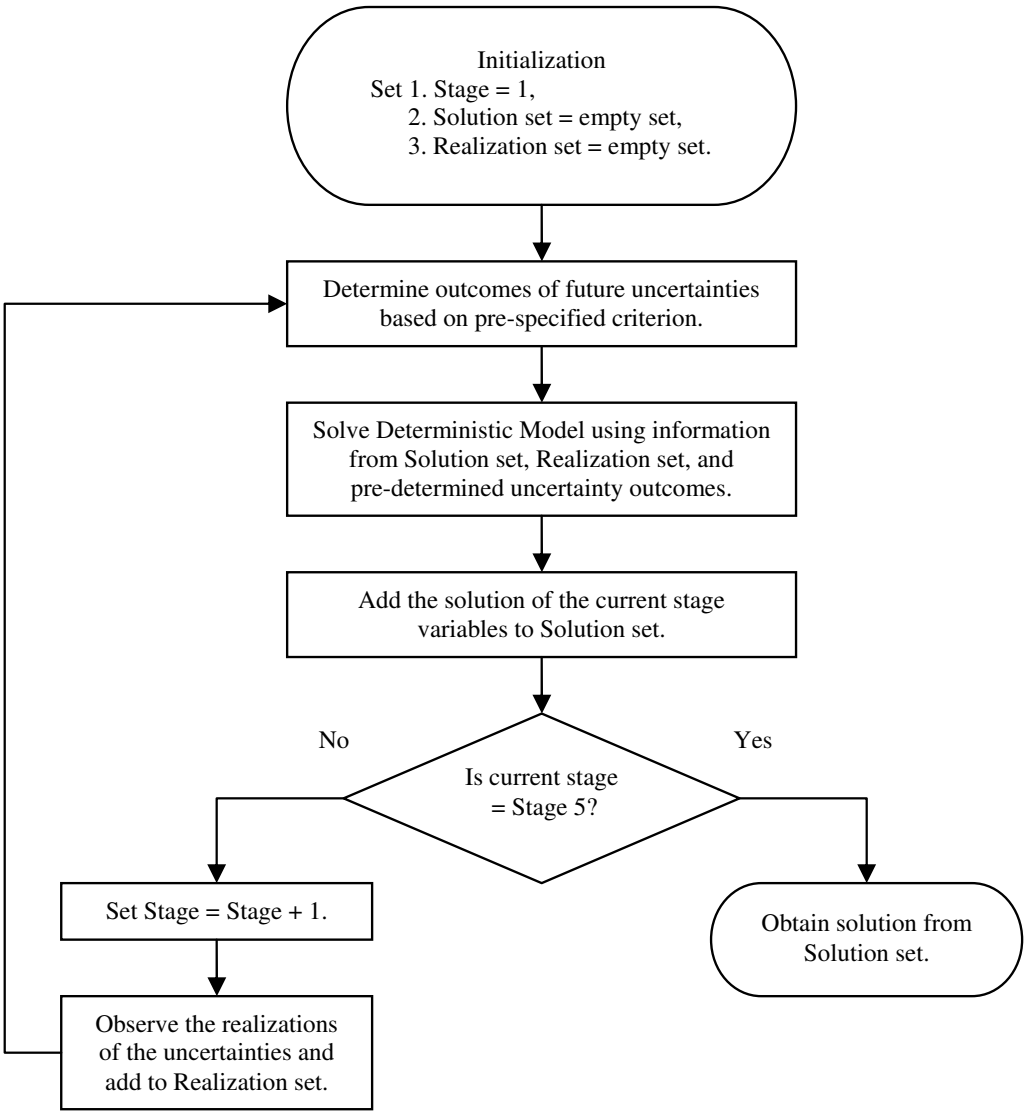

Figure 35: Flow chart of the heuristic approaches based on greedy algorithms

1. Greedy Deterministic Approach **(GDA)** - this approach chooses the uncertainty outcome that is *most likely* to occur as a representative of the future uncertainty.

2. Greedy Optimistic Deterministic Approach **(GOA)** - this approach chooses the *best* outcome as a representative of the future uncertainty.

3. Greedy Pessimistic Deterministic Approach **(GPA)** - this approach chooses the *worst* outcome as a representative of the future uncertainty.

GDA considers outcomes of the upcoming uncertainties as ones that have the highest chance to occur, which represents the strategy of the risk-neutral farmer. GOA always determines the future uncertainties in the optimistic way. This approach may represent the strategy of the risk-taking farmer. On the other hand, GPA considers in pessimistic way and may represent the strategy of the risk-averse farmer.

4.6 Numerical Study

In this section, we represent the application of the stochastic programming and heuristics described in Sections 4.4 and 4.5. We consider the case of a representative farmer who grows corn and soybean in Illinois in 2005. Corn and soybean are chosen because they account for about 80% of the total grain and oilseed production in the U.S. In addition, Illinois is ranked first in producing soybean and second in corn. According to Acreage Report (National Agriculture Statistics Service 2005), 93% of the planted area in Illinois is devoted to corn and soybean.

4.6.1 Data Background

The planning periods span from the first planting date to the end of the year. According to the Usual Planting and Harvesting Dates for U.S. Field Crops report (National Agricultural Statistics Service 1997), in Illinois, corn is planted from the end of April to the end of May while soybean is from the beginning of May to the beginning of June. Likewise, corn is harvested from the end of September to the middle of

November and soybean from the end of September to the beginning of November. In addition, we define the after-harvest periods as the periods from the end of the harvest season to the end of year. We define the time periods in the model on a weekly basis. The first week of the planting period is marked as week or period 1 and the last week of the year is set as week or period 35. Under this setting, the planting periods, harvesting periods, and after-harvest periods correspond to weeks 1 to 6, 22 to 29, and 27 to 35, respectively. Note that the overlap in harvesting periods and after-harvest periods is due to the inconsistency between harvesting periods of corn and soybean.

The data for the uncertainties, yields and prices, are derived from the yield and price forecastings in Chapters 2 and 3, respectively. We assume that each of these uncertainties has three levels - high, medium, and low. These three levels of yield1 correspond to the quantiles 0.975, 0.500, and 0.025 of the yield prediction band, respectively. On the other hand, these three levels of yield2 correspond to 100%, 90%, and 80% of yield1, respectively. Similar to yield1, the high, medium, and low levels of price_h1, price_h2, and price_e are the quantiles 0.975, 0.500, and 0.025 at the middle of the harvest season, the end of the harvest season, and the end of year of the price prediction band, respectively. Table 7 shows the forecasts of corn yields and prices while Table 8 shows the forecasts for soybean. The prior probabilities associated with the uncertainty outcomes are presented in Table 22 in Appendix B. The prior probabilities reflect the farmer's belief on the outcomes of yields and prices that will occur this year. They may depend on available information, experience, and the farmer's risk characteristic.

Resource parameters, land and labor, are set according to 2002 Census of Agriculture (National Agricultural Statistics Service 2004). Most of the farms in Illinois have a single operator. Therefore, we assume that our representative farm has one

Table 7: Forecasts of corn yields and prices in 2005

Uncertainty	High	Medium	Low
yield1 (bushel/acre)	209.66	170.49	131.32
yield2 (bushel/acre)	100% of yield1	90% of yield1	80% of yield1
price_h1 ($/acre)	2.0741	2.0229	1.9717
price_h2 ($/acre)	2.1803	2.1339	2.0875
price_e ($/acre)	2.2041	2.1619	2.1197

Table 8: Forecasts of soybean yields and prices in 2005

Uncertainty	High	Medium	Low
yield1 (bushel/acre)	55.42	46.68	37.94
yield2 (bushel/acre)	100% of yield1	90% of yield1	80% of yield1
price_h1 ($/bushel)	6.0419	5.9549	5.8679
price_h2 ($/bushel)	6.0966	5.9962	5.8958
price_e ($/bushel)	6.2269	6.1446	6.0623

operator. We set the available labor hours for each week during planting and harvesting periods at 40 hours, i.e. a farmer works 8 hours a day, five days a week. The available cultivation area is assumed to be 330 acres, which is equal to the weighted average of farm sizes with one operator in Illinois. Cost parameters, in contrast, are approximated as follows. We estimate planting and harvesting costs from Grain Farm Returns and Costs, Illinois 2006 report (University of Illinois 2005). We adjust by subtracting storage and hired labor costs from the total non-land costs of Central Illinois - High Productivity Farmland since we determine storage cost as another parameter and no farm workers are hired. We assume that 75% of the adjusted cost goes to planting cost and the remaining 25% is harvesting cost. Planting and harvesting costs of corn and soybean are shown in Table 9.

Storage cost is set according to the Prairie Grain magazine (Minnesota Association of Wheat Growers, North Dakota Grain Growers Association and South Dakota Wheat, Inc. 2005). On-farm storage costs are $0.00325/bushel/week for corn and

Table 9: Planting and harvesting costs of corn and soybean in 2005

Crop	Planting Cost ($/acre)	Harvesting Cost ($/acre)
Corn	201.75	139.75
Soybean	67.25	43.25

Source: [114]

$0.00975/bushel/week for soybean. For transportation cost, we use information from The Leasing Forum newsletter (University of Illinois 2001) which is $0.45/bushel for both corn and soybean. Labor used in planting and harvesting operations is acquired from MU Guide (University of Missouri-Columbia 1997). Labor hours needed for planting one acre of corn and soybean are fixed at 0.81 hour and 0.71 hour, respectively. In contrast, we assume that labor hours required to harvest an acre of corn and soybean during the first and second halves of the harvest season depend on the realizations of the yield1 and yield2, respectively. We employ the interest rate quoted by Commodity Credit Corporation, which is 5.875% per year or 0.1123% per week. In addition, there is a minimum requirement amount for corn at the end of the harvest season. It is set to 8,000 bushels which is needed for cattle feed. Finally, the cost or penalty to purchase corn from the market to satisfy the minimum requirement amount is set to 35% more than the selling price at that time. Parameters described so far, except yield, price, planting cost, harvesting cost, and labor hour needed for harvesting, are summarized in Table 10. Table 11 shows the labor hour needed for harvesting an acre of corn and soybean.

Next, we consider a small example where corn and soybean are assumed to have the same uncertainty outcomes in every stage. We relax this assumption in a medium example where corn and soybean can have different uncertainty outcomes.

Table 10: Parameter values used in the model (except yield, price, planting cost, harvesting cost, and labor hour needed for harvesting)

Parameter	Value	Unit
Land	330	acre
Available labor hours during planting periods	40	hour/week
Available labor hours during harvesting periods	40	hour/week
Storage cost for corn	0.00325	$/bushel/week
Storage cost for soybean	0.00975	$/bushel/week
Transportation cost for corn	0.45	$/bushel
Transportation cost for soybean	0.45	$/bushel
Labor hour needed for planting corn	0.81	hour/acre
Labor hour needed for planting soybean	0.71	hour/acre
Interest rate	0.1123	%/week
Minimum requirement for corn	8000	bushel
Penalty to purchase corn from the market [defined as % of selling price at that time]	35	%

Sources: [20], [74], [82], [115], [116]

Table 11: Labor hour needed for harvesting corn and soybean

Parameter	High	Medium	Low
Labor hour needed for harvesting corn during the first half of the harvest season (hour/acre)	0.42	0.32	0.22
Labor hour needed for harvesting soybean during the first half of the harvest season (hour/acre)	0.42	0.32	0.22
Labor hour needed for harvesting corn during the second half of the harvest season (hour/acre) [defined as % of labor needed during the first half]	100%	90%	80%
Labor hour needed for harvest soybean during the second half of the harvest season (hour/acre) [defined as % of labor needed during the first half]	100%	90%	80%

Source: [116]

Table 12: Planting area (in acre) during planting periods of the small example

Period	SLP		GDA		GOA		GPA	
	Corn	Soybean	Corn	Soybean	Corn	Soybean	Corn	Soybean
1	49.383		48.310		49.383		49.383	
2		56.338		56.338	49.383			56.338
3		56.338		56.338	49.383			56.338
4	7.617	47.648		56.338	49.383		11.537	43.176
5		56.338		56.338		56.338		56.338
6		56.338		56.338		56.338		56.338
Subtotal	57.000	273.000	48.310	281.690	197.532	112.676	60.920	268.528
Total	**330.000**		**330.000**		**310.208**		**329.448**	

4.6.2 Small Example

We illustrate the application of stochastic programming and heuristics using the data described in Section 4.6.1. In this example, we assume that corn and soybean have the same uncertainty outcomes. Since there are five uncertainties and each of them has three possible outcomes, there will be $3^5 = 243$ scenarios in total.

One of the most important decisions that the farmer makes is the acreage allocation which is the first-stage decision. This is because the grower has to make this decision without the certain information about yields and prices. In addition, this is an irreversible decision since the farmer cannot grow more crops that actually give higher returns at the time when the yields and prices are realized. The acreage allocation, accompanied with the planting scheduling, is indirectly presented by the planting area during planting periods in Table 12.

From Table 12, results from SLP indicate that the farmer should cultivate on all available land where most of the land is allocated to soybean. Only 17% of the land is allocated to corn. This may imply that in 2005, soybean provided a higher return per acre than corn did. Corn is mostly planted in period 1 because it is not the soybean planting period. The reason that SLP allocates some resources to grow corn in period 4 rather than use all resources to grow soybean comes from the minimum

requirement amount of corn at the end of the harvest season. This approach focuses on minimizing the amount of corn purchased when the corn yield is low as well as the use of available land. In addition, corn is grown in period 4 rather than period 2 because of the time value of money (TVM). TVM is accounted by the compounding factors in equation (20). The time value of money means that under positive interest rate environment, a person prefers to receive a certain amount of money today, rather than the same amount in the future. The reason is that money received today is more valuable than money received in the future since it can earn interest from depositing in the bank or it can be re-invested. On the other hand, one prefers to pay money in the future, rather than the same amount of money today from the same reasons. Since corn has higher planting cost than soybean, the farmer should grow corn as late as possible which is period 4, the last planting period for corn.

Results from GDA also suggest using all of the land to grow crops. Similar to SLP, most of the land is allocated to soybean. In fact, the farmer grows soybean as much as possible. The reason that the farmer does not plant only soybean comes from the limitation of the available planting labor during the soybean planting periods. Hence, he can grow at most 56.338 acres a week. For the rest of the available land, he grows corn. These decisions show the shortsighted strategy of this heuristic compared to SLP. Since it considers only the outcome that is most likely to occur for the yield uncertainty, it ignores the possibility that corn may have a low yield. If the low corn yield is realized, the farmer has to buy a large amount of corn at the penalty price to satisfy the minimum requirement. This will lower the profit he will receive.

In contrast to GDA, results from GOA recommend the farmer to grow corn as much as possible. He grows corn 49.383 acres per week every week during corn planting periods. The rest of the land is allocated to soybean. Since he can grow soybean only 56.338 acres per week and he has only 2 weeks left, he cannot use all of his land under this strategy. Given that this heuristic assumes the best outcome

as a representative of each uncertainty, this may imply that when yields and prices are at their high level, corn will give a higher profit than soybean. This heuristic will not have a difficulty to satisfy the minimum requirement amount when the corn yield is low. This is because the farmer grows corn on most of his land. However, when yields and/or prices are not at the high level, he may obtain less profit since on the average, soybean generated a higher return than corn in 2005 as determined by SLP.

Results from GPA have the planting decision very similar to the one from SLP. Most of the land is allocated to soybean. Corn is cultivated only in periods 1 and 4. However, this heuristic suggests to grow corn in period 4 more than stochastic programming model does. This is because GPA looks at the uncertainties in a pessimistic way. It assumes that crops will certainly have low yields. Therefore, the farmer should grow enough corn to satisfy the minimum requirement amount when the corn yield is low. Because corn requires more planting labor than soybean, it makes the total cultivated area less than the available land.

Other decisions from stochastic programming and heuristic approaches are summarized as follows.

In the strategy determined by SLP, all corn is harvested in period 24, regardless of the outcome of yield1. Soybean is harvested as much as and as late as possible in the first half of the harvest season to avoid the possibility of yield loss in the second half (yield2) and to pay less storage cost. The farmer can harvest all soybeans in the first half of the harvest season when yield1 is either medium or low. However, if yield1 is high, 44.286 acres of soybean will be harvested in the second half. At the middle of the harvest season, no corn is sold even though corn price (price_h1) is high. On the other hand, all soybean available up to this time is sold only if the soybean price is high. The surplus of corn, after allocating for the minimum requirement, is sold at the end of the harvest season when its price (price_h2) is either high or medium. However, if yield1 is low, the farmer will not have enough corn to satisfy the requirement so he

cannot sell corn but he has to purchase the shortage, 515 bushels, from the market. The remaining of soybean, if not sold at the middle of the harvest season, will be sold if price_h2 for soybean is high. If there are crops stored at the end of the year, they will be sold to the market regardless of the price at that time (price_e).

Since heuristics generate only the decision rule for a particular scenario, we further enumerate all 243 possible scenarios to generate the complete heuristics' policies to compare with the policy from stochastic programming approach.

Under the strategy from GDA, the farmer does not harvest corn during the first half of the harvest season but harvests some soybean during these periods if yield1 is either medium or low and the rest during the second half. As mentioned above, this strategy considers only the most likely outcomes hence it assumes that yield2 of corn will be the same as yield1. All corn is harvested at the end of the harvest season, regardless of yield1, yield2, and price_h1. Soybean is sold at the middle of the harvest season if the price_h1 is high and yield2 is either high or medium or if the price_h1 is medium and yield2 is high. Since the farmer grows corn only 48 acres and also harvests at the end of the harvest season, he has to purchase corn from the market when yield1 is either medium or low. He can produce enough corn for the minimum requirement only when the corn yiled1 is high. He sells all of his surplus corn, corn produced minus the minimum requirement, at the end of the harvest season, regardless of the price_h2 of corn at that time. Soybean is sold only if its price_h2 is high. Finally, any soybean left in the storage will be sold at the end of year.

GOA determines the uncertainties only on their best outcomes. Hence, it assumes that yield2 will equal yield1 and prices will be at their high levels. The farmer using this strategy does not harvest any crop during the first half of the harvest season and also not sell at the middle of the harvest season. He schedules to harvest his crops as late as possible in the second half with higher priority on corn. This is because corn has higher harvesting cost than soybean so it is better to harvest corn after soybean.

Since most of his land is allocated to corn, he can satisfy the minimum requirement for any outcomes of yield1 and yield2. He sells his crops at the end of the harvest season only if their prices (price_h2) are high. Otherwise, he will keep them and sell at the end of year.

The farmer who uses the strategy from GPA schedules to harvest all corn at the middle of the harvest season, to avoid the lower yield after that period. He also harvests soybean as much as possible using the remaining labor hours after allocating to harvest corn. This is because this strategy looks at the worst outcomes of the uncertainties then it assumes that yield2 will be much lower than yield1 so it is better to harvest within the first half of the harvest season. However, 43.734 acres of soybean are harvested at the second half because the farmer does not have enough labor to harvest all of them in the first half. The farmer sells corn left after reserving for the minimum requirement at the middle of the harvest season if price_h1 is high. On the other hand, he sells soybean available on hand when price_h1 is high or medium. Since this strategy can produce corn at least 8,000 bushels when yield1 is low (60.92 acres $\times$ 131.32 bushels/acre), no corn will be bought from the market. All corn in the storage at the end of the harvest season is sold at this time, regardless of the price (price_h2). At the same time, soybean is sold only if price_h2 for soybean is high or medium. Otherwise, it will be sold at the end of year.

In order to evaluate the performance of the proposed approaches, we randomly generate 10,000 instances where each uncertainty has a chance to occur according to the prior probabilities specified in Table 22 in Appendix B. The numerical results of the stochastic programming model and heuristics in terms of expected (or long run average) profit and average time in seconds used to generate the policy or solution are presented in Table 13. The computation is performed on Pentium M 1.6 GHz with 1 GB RAM using CPLEX 10 as a solver in GAMS 22.4.

It is not surprising that SLP gives the highest expected profit since, by definition,

Table 13: Performance comparison of the small example

Performance	SLP	GDA	GOA	GPA
Average profit ($)	12267.75	7797.29	2532.87	11917.14
% of optimal profit	100.00	63.56	20.65	97.14
Average time (second)	0.56	0.68	0.69	0.70

it delivers the policy that maximizes the expected profit. The farmer makes $12,268 profit from the SLP's strategy. Hence, $12,268 is the optimal solution of the small example.

GDA generates $7,797 which is 63.56% from the optimal solution. The main reason that this strategy produces less than the optimal solution is that under this policy, the farmer does not grow enough corn to satisfy the minimum requirement amount when yield1 is medium or low. Therefore, he has to pay the penalty for purchasing corn from the market. In addition, under this policy, he harvests most of his crops in the second half of the harvest season which may lead to have a yield loss.

The smallest expected profit comes from GOA. GOA provides only $2,533, about one-fifth of the optimal solution. The farmer grows only 310 acres out of 330 acres so he gives up the opportunity to generate profit on the remaining area. Moreover, he harvests all crops in the second half of the harvest season. Hence, when the yield2 is medium or low, he loses large portion of his produces and this loss results in very low profit under these scenarios. Besides, this strategy allocates most of the land to corn planting which generally provides less return than soybean in 2005.

Finally, GPA generates $11,917 which equals to 97.14% of the optimal solution. This is because GPA has a very similar strategy as one from SLP. Both methods have almost the same acreage allocation between corn and soybean. They also use the same criteria for harvesting scheduling during the first and second halves of the harvest season. However, since GPA considers only worst outcomes for the future uncertainties, its selling decisions are different from SLP. Consequently, GPA provides

the less-than-optimal profit.

In terms of the average computational time, the stochastic programming model can solve the problem slightly faster than the heuristic approaches. The reason is that this is a small example with only 243 scenarios. Therefore, stochastic programming model can be solved in a very short time. All three heuristics, on the other hand, are based on greedy algorithms which have to re-solve the problem in every stage. Since there are five stages, these heuristics have to re-solve the problem five times. Consequently, the heuristic approaches solve the problem slower than stochastic programming approach. However, this result will not be the same when the problem size is large. This will be illustrated with the following example.

4.6.3 Medium Example

In this example, corn and soybean may have different uncertainty outcomes. However, we assume that outcomes of corn and soybean are independent. Hence, in each uncertainty, there will be 9 possible outcomes. The number of scenarios increases to $9^5 = 59{,}049$ scenarios. Since both crops are independent, the (conditional) probability will be the product of the (conditional) probability of corn and the (conditional) probability of soybean, i.e.

$$\begin{aligned} P(s_1 \text{ of corn} = High \text{ and } s_1 \text{ of soybean} = High) &= P(s_1 \text{ of corn} = High) \times \\ & \quad P(s_1 \text{ of soybean} = High). \end{aligned}$$

We assume that both crops have the same prior probabilities as defined in Table 22 in Appendix B. In this example, we investigate only the acreage allocation and the performance among the approaches we proposed. The acreage allocation and the corresponding planting scheduling are shown in Table 14. The allocation decision from the medium example is exactly the same as the decision from the small example. This is because both crops are assigned the same prior probabilities as used in the small

Table 14: Planting area (in acre) during planting periods of the medium example

Period	**SLP**		**GDA**		**GOA**		**GPA**	
	Corn	Soybean	Corn	Soybean	Corn	Soybean	Corn	Soybean
1	49.383		48.310		49.383		49.383	
2		56.338		56.338	49.383			56.338
3		56.338		56.338	49.383			56.338
4	7.617	47.648		56.338	49.383		11.537	43.176
5		56.338		56.338		56.338		56.338
6		56.338		56.338		56.338		56.338
Subtotal	57.000	273.000	48.310	281.690	197.532	112.676	60.920	268.528
Total	**330.000**		**330.000**		**310.208**		**329.448**	

example. The explanation for the first stage decision of each approach is therefore the same as we describe in the small example.

Since the medium example has the same first-stage decision as the small example, even though they have different number of scenarios, we can expect that they should have similar long run average profit. This conjecture is confirmed by the expected profits of the medium example in Table 15. We run the simulation by randomly generating 60,000 instances and compute the long run average profit and the average time required to solve the problem. The computation is performed on the same setting as in the small example. The average profits are similar to the previous example except for the case of GDA. This is because the small and medium examples have different predetermined uncertainty outcomes. Therefore, the decisions in later stages are not the same and the average profits are different. On the other hand, GOA (GPA) always chooses the best (worst) outcome as the representative of uncertainty, which coincides in both example, i.e. yield1 is high (low) for both corn and soybean. Consequently, the decisions are similar in the small and medium examples. Overall, SLP yields the optimal and hence the highest average profit. GOA gives the lowest average profit while GPA delivers the near optimal solution.

For the average computational time, in this example, the stochastic programming

Table 15: Performance comparison of the medium example

Performance	SLP	GDA	GOA	GPA
Average profit ($)	12337.22	5853.92	2749.00	12178.07
% of optimal profit	100.00	47.45	22.82	98.71
Average time (second)	30.30	1.58	1.50	1.40

model requires more time to generate the policy, about 50 times that of the small example, while the heuristic models use about twice the time used in the small example. The reason behind this is that the number of scenarios increases dramatically in the medium example. Hence, stochastic programming takes a much longer time to consider every node to determine the optimal solution. On the other hand, the heuristic approaches still re-solve five times as in the small example so they can solve the problem in a reasonable amount of time. The difference in the computational times between stochastic programming and heuristics will be amplified when the problem is larger than the example we illustrate here.

We can conclude that our heuristics are more suitable, in terms of computational time, to solve the decision planning under uncertainty than stochastic programming, especially in the very large case. The profits generated by the heuristic approaches are reasonable. GPA provides the near optimal solution.

Next, we investigate the robustness of the model to the misspecifications of the information we use in this study.

4.6.4 Sensitivity Analysis

In this section, we investigate the robustness of the results from our model to model misspecifications. First, we examine the robustness to yield and price uncertainties and then we explore the robustness to the prior probabilities.

4.6.4.1 Robustness to Uncertainties

Since the yields and prices are regarded as uncertainties in the decision planning models and their numerical data are estimated from the forecasting methods, it is necessary to evaluate the robustness of the model results to their estimates. In our problem, we have five uncertainties, two from yields (yield1 and yield2) and three from prices (price_h1, price_h2, and price_e). However, yield2 is defined as the percentage of yield1, therefore changing the value of yield1 will proportionally change the value of yield2. Hence, we consider only four uncertainties. We denote yield1 as factor A, price_h1 as factor B, price_h2 as factor C, and price_e as factor D.

We design the experiment to study the robustness of our model to these factors. We use a factorial design that is most efficient for our experiment (Montgomery 1997). We determine all possible combinations of the factor levels we are interested. In our study, each factor has two levels. We refer to these levels as Max and Min. Consequently, we require $2 \times 2 \times 2 \times 2 = 16$ cases. This design is a particular type of a 2^k factorial design, where $k = 4$ is the number of factors. We run only one replication for each case, since the results from the optimization models do not vary from one replication to another.

The Max and Min are derived as follows. We calculate Max and Min of yield1 from the yield forecasting Model 3 described in Section 2.4.3. First, we detrend the forecast of the yield by removing the contribution from GDP. Second, the maximum and minimum of the detrended yield values from 1927 to 2004 are obtained. Third, we add the contribution of GDP in 2005 back to these selected values to obtain the Max and Min of yield1 in 2005. Recall that, in this study, we assume yield1 has 3 levels - high, medium, and low. The Max and Min values will be used as the medium level for yield1. The differences between the levels are still the same as we use in the small and medium examples, 39.17 bushels for corn and 8.74 bushels for soybean.

The cash or selling prices (price_h1, price_h2, and price_e) are computed by adding

Table 16: Values of the factors of corn used in the factorial design

Factor	Description	Max Level			Min Level		
		High	Medium	Low	High	Medium	Low
A	yield1 (bushel/acre)	236.37	197.20	158.03	181.29	142.12	102.95
B	price_h1 ($/bushel)	2.5091	2.4579	2.4067	1.8291	1.7779	1.7267
C	price_h2 ($/bushel)	2.6153	2.5689	2.5225	1.8528	1.8064	1.7600
D	price_e ($/bushel)	2.5966	2.5544	2.5122	1.8666	1.8244	1.7822

the forecasted commodity basis to the futures prices as described in Section 3.3. Since the commodity basis is very small compared to the futures prices and the futures prices change continuously, we calculate the Max and Min of the cash prices from the highest and lowest values of the corresponding futures prices. We find the highest and lowest values of the futures prices within one year before the time periods of the corresponding cash prices. Then we add the commodity basis forecasts to the selected futures prices to get the Max and Min of the cash prices. For example, let price_e be the corn price at the end of 2005. The corresponding corn futures contract is March 2006. We search the maximum and minimum of March 2006 corn futures prices from the end of 2004 to the end of 2005, one year before time period of price_e (the end of 2005). Recall that a futures contract is traded several years in advance. Figure 36 shows the March 2006 corn futures contract prices from the end of 2004 to the end of 2005. Similar to yield1, these cash prices have three levels and the Max and Min will be used as the medium level of the prices. The deviations between levels are the same as we use in the examples. The Max and Min levels of the factors of corn and soybean are displayed in Table 16 and Table 17, respectively.

We solve the stochastic programming and heuristic models using the Max and Min information. We use symbols "+" for Max and "-" for Min. The results from the factorial design are presented in Table 18 for the small example and Table 19 for the medium example. As we discuss in Section 4.6.3, the results from the small

Table 17: Values of the factors of soybean used in the factorial design

Factor	Description	Max Level			Min Level		
		High	Medium	Low	High	Medium	Low
A	yield1 (bushel/acre)	61.64	52.90	44.16	49.57	40.83	32.09
B	price_h1 ($/bushel)	7.4794	7.3924	7.3054	5.0294	4.9424	4.8554
C	price_h2 ($/bushel)	7.5541	7.4537	7.3533	5.1291	5.0287	4.9283
D	price_e ($/bushel)	7.6844	7.6021	7.5198	5.2594	5.1771	5.0948

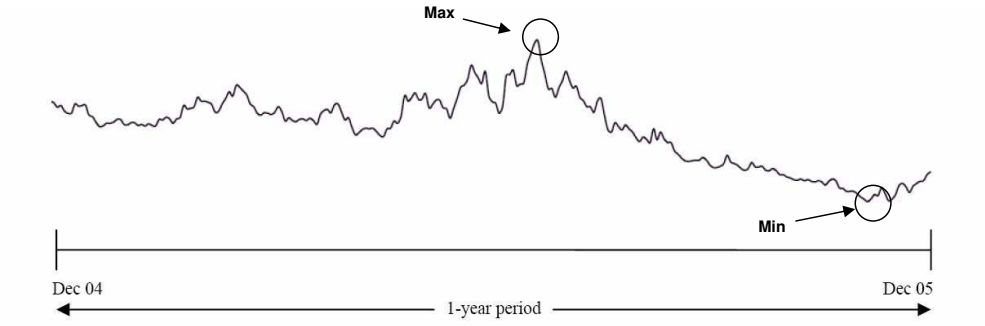

Figure 36: March 2006 corn futures contract prices from December 2004 to December 2005

and medium examples are not much different, except for GDA which has moderate differences in some cases. Overall, all approaches will make high returns when yield1 is at its high level and low returns when yield1 is at its low level. Case 1 is the only case that has a loss. This is because every factor is at its low level, i.e. low yield and low prices. A small profit is generated in Case 2, where yield1 is high and all prices are low. Other than these two cases, when every factor remains the same except for the price, the profits do not change much. Therefore, *we may conclude that the stochastic programming and heuristic models are robust to the prices but not to the yield.*

Table 18: Factorial design result of the small example

Case	**Factor**				**Average Profit ($)**			
	A	B	C	D	**SLP**	**GDA**	**GOA**	**GPA**
1	-	-	-	-	-11899.06	-16273.17	-16572.75	-19602.07
2	+	-	-	-	10467.78	4907.44	1932.02	10384.82
3	-	+	-	-	15593.88	12776.92	12070.82	12439.14
4	+	+	-	-	47273.38	43605.39	42437.92	46854.71
5	-	-	+	-	14907.72	8062.18	8006.50	14276.26
6	+	-	+	-	47658.22	38996.63	36962.42	47543.96
7	-	+	+	-	15148.43	10088.84	8004.48	14191.54
8	+	+	+	-	47895.38	41925.98	36933.34	47678.87
9	-	-	-	+	16005.01	9147.52	9433.42	12811.51
10	+	-	-	+	47829.62	39076.18	35568.68	47466.96
11	-	+	-	+	16225.34	12411.48	9360.03	12786.43
12	+	+	-	+	48108.05	42844.67	35590.67	47605.94
13	-	-	+	+	15109.95	8169.06	8364.59	14554.95
14	+	-	+	+	47986.85	39322.33	36823.91	47814.92
15	-	+	+	+	15076.73	10491.70	8272.00	14407.29
16	+	+	+	+	48103.07	41961.49	36692.79	47604.01

Table 19: Factorial design result of the medium example

Case	Factor				Average Profit ($)			
	A	B	C	D	**SLP**	**GDA**	**GOA**	**GPA**
1	-	-	-	-	-11558.27	-16287.82	-16476.68	-19688.32
2	+	-	-	-	10752.11	4927.40	2036.78	10378.34
3	-	+	-	-	15568.57	12067.08	12194.17	12203.81
4	+	+	-	-	47123.38	42628.62	42715.44	46884.15
5	-	-	+	-	15024.11	8240.80	8316.10	14484.20
6	+	-	+	-	48010.02	39341.69	37379.51	47795.43
7	-	+	+	-	15136.58	8571.16	8262.31	14391.61
8	+	+	+	-	48134.19	39323.87	37358.08	47622.15
9	-	-	-	+	16215.49	9428.63	9393.22	12755.41
10	+	-	-	+	48052.59	39346.78	35817.43	47873.76
11	-	+	-	+	16220.75	9580.90	9350.88	12614.76
12	+	+	-	+	48146.88	39235.15	35573.53	47524.58
13	-	-	+	+	15295.26	8506.19	8283.64	14861.99
14	+	-	+	+	48141.16	39697.16	37250.04	47836.23
15	-	+	+	+	15447.16	8500.22	8490.92	14620.29
16	+	+	+	+	48310.56	39498.11	37173.31	47553.63

Next, we employ the Analysis of Variance (ANOVA) technique to formally examine the statistical significance of the factors. We assume that there is no third-order interaction.

Define α as a Type I error, that is the error when we reject the null hypothesis when the null hypothesis is true. We will provide the results at $\alpha = 0.05$ and 0.10. ANOVAs for the small and medium examples are provided in Appendix C. The observations from the ANOVAs can be summarized as follows.

- **SLP**

  1. ANOVAs for the small and medium examples have the same factors that are significant at $\alpha = 0.05$ and $\alpha = 0.10$.
  2. Only yield1 is significant at $\alpha = 0.05$.

3. Prices and the interactions among prices (first- and second-order) are significant at $\alpha = 0.10$.

- **GDA**

1. ANOVAs from the small and medium examples have different factors that are significant at $\alpha = 0.10$.
2. Price_h1, price_e, and the interactions that have these prices are significant at $\alpha = 0.10$ in the small example.
3. Price_h1 and the interactions that have price_h1 are significant at $\alpha = 0.10$ in the medium example.
4. Both examples have only yield1 that is significant at $\alpha = 0.05$.

- **GOA**

1. Only yield1 is significant at $\alpha = 0.05$ in both examples.
2. No other factor or interaction is significant at $\alpha = 0.10$.

- **GPA**

1. ANOVAs of GPA are similar to ones from SLP in terms of the significant factors.
2. ANOVAs for the small and medium examples have the same factors that are significant at $\alpha = 0.05$ and $\alpha = 0.10$.
3. Only yield1 is significant at $\alpha = 0.05$.
4. Prices and the interactions among prices (first- and second-order) are significant at $\alpha = 0.10$.

We can formally conclude that yield1 is the only factor that is significant at $\alpha = 0.05$ and in general, the prices are also significant but less than yield1. *These*

ANOVA results are consistent with the conclusion from the factorial design. This is because the prices are less significant than yields. Consequently, the proposed models are robust to the prices more than to the yield. With this result, we should pay more attention in the yield data that we use in the decision planning models.

4.6.4.2 Robustness to Prior Probabilities

In our numerical study, we use pre-specified or prior probabilities for the outcomes of each uncertainty. These probabilities are subjective and depend on the decision makers' point of view. Hence, the prior probabilities may not be the same for everyone. Consequently, it is necessary to examine the robustness of the model results to the prior probabilities. In this section, we investigate the robustness of the objective values to the prior probabilities by using the different sets of prior probabilities.

First, we adjust only the probabilities for yield1 and yield2 to have equal chance for each outcome to occur. Next, we assign equal chance to all uncertainties. The adjusted prior probabilities are shown in Tables 23 and 24 in Appendix B.

Note that, by assigning equal probability to the uncertainties, GDA will choose the outcomes randomly since they have the same chance to occur and GDA selects the outcomes that are most likely to occur. Hence, we discard GDA from this investigation.

The average profits from SLP, GOA, and GPA under all prior probability settings are illustrated in Table 20 for the small example and Table 21 for the medium example.

The average profits from the small and medium examples are similar. Overall, SLP and GPA yield about the same level of expected profits under all prior probability settings. For GOA, the expected profits under adjusted settings are negative while the expected profit under the original setting is small but positive. Hence, *we can conclude that SLP and GPA approaches are robust to the pre-specified probabilities but GOA is not.*

Table 20: Expected profit comparison under different prior probability settings of small example

Prior Probabilities	**SLP**		**GOA**		**GPA**	
	Profit ($)	% of optimal	Profit ($)	% of optimal	Profit ($)	% of optimal
Original setting	12267.75	100.00	2532.87	20.65	11917.14	97.14
Adjusted setting with equal chance for yields	12039.00	100.00	(436.89)	(3.63)	11650.05	96.77
Adjusted setting with equal chance for yields and prices	12175.74	100.00	(298.72)	(2.45)	11876.80	97.54

Table 21: Expected profit comparison under different prior probability settings of medium example

Prior Probabilities	**SLP**		**GOA**		**GPA**	
	Profit ($)	% of optimal	Profit ($)	% of optimal	Profit ($)	% of optimal
Original setting	12337.22	100.00	2749.00	22.28	12178.07	98.71
Adjusted setting with equal chance for yields	12141.16	100.00	(579.51)	(4.77)	11801.12	97.20
Adjusted setting with equal chance for yields and prices	12243.58	100.00	(92.15)	(0.75)	11923.18	97.38

4.7 Conclusions

In this chapter, we develop the crop decision planning models that help farmer make decisions when there are uncertainties in yields and prices. The main objective of the planning models is to maximize the expected profit. We propose a model under stochastic programming framework, which yields the optimal expected profit. However, the stochastic programming approach itself has a limitation when it is used to solve large size problem since the computational effort increases exponentially with the number of decision stages. Hence, we propose three heuristics based on greedy algorithms to solve the same problem, which yield feasible solutions within polynomial times. We make use of the forecasted yields and prices from Chapters 2 and 3, respectively. Even though the yields and prices are forecasted as functionals, we assume discrete outcomes where each of these forecasts has three levels - high, medium, and low. A numerical study is performed to validate both stochastic programming and heuristic models in the small and medium examples. The results from the proposed approaches are compared in terms of the average profits and average computational time. The sensitivity analysis is also investigated. A factorial design is employed to examine the robustness of the model results to the uncertainty factors. Finally, the average profits under different prior probability settings are determined to evaluate the robustness of model results to the pre-specified probabilities.

The numerical study shows that the acreage allocation, the first-stage decision, is exactly the same in the small example where corn and soybean are assumed to have the same uncertainty outcomes and in the medium example where both crops are independent but have the same prior probabilities. SLP which yields the optimal solution gives the highest weight to grow soybean and to the use of available land. Corn is grown just to minimize the purchased amount in order to satisfy the corn minimum requirement constraint. GPA has similar acreage allocation decision as SLP. However, this approach gives the highest priority on corn minimum requirement. It

guarantees that farmer will not have to purchase corn when the corn yield is low. GDA gives the highest priority on soybean. Soybean is planted as much as possible and the rest of land is then allocated to corn. In contrast, corn is grown as much as possible under GOA strategy. *In terms of the expected profits, GPA generates high profit, about 97%-99% of the optimal solution from SLP. GDA provides a moderate return, about 47%-64% of the optimal solution. GOA, on the other hand, gives the lowest profit, only 21%-23% of the optimal solution.*

The benefit of heuristic approaches is prominent when the problem size is large. The average time used by the heuristic approaches to solve the problem is much less than the average time used by the stochastic programming model. Since the heuristic approaches provide feasible solutions in the reasonable amount of time, especially GPA that gives the near optimal expected profit, it may be better to use the proposed heuristics to solve the large-scale problem rather than the stochastic programming. However, if the problem size is not too large, stochastic programming is a good alternative since it generates the complete policy and also the optimal solution.

According to the sensitivity analysis, a factorial design of experiment reveals that *all uncertainties have significant effects on the average profits but yield has higher impact than prices. In addition, the proposed models are robust to the prices but not to the yield.* We also find that *SLP and GPA are robust to the prior probabilities but GOA is not.*

CHAPTER 5

CONCLUSIONS AND FUTURE RESEARCH

5.1 Summary

This dissertation develops a crop decision planning model under uncertainty. More specifically, it focuses on the farm-level decision planning for farmers who grow storable crops under yield and price uncertainties. We develop forecasting models to predict the possible outcomes of the uncertainties, which are further integrated in the crop decision planning model. In particular, we consider the case of a representative farmer who grows corn and soybean in Illinois.

In Chapter 2, we present a crop-weather regression model to predict the crop yield. The weather information used in the forecasting model are temperature and rainfall during the growing season. We also incorporate the GDP information to account for the economic growth, which indirectly reflects in the yield change over time. Unlike most of the regression models that use a parametric framework, we develop a semiparametric regression model that determines the within- and between-year relationships for the data. We use the concept of functional principal component analysis (FPCA) to reduce the dimensionality of the model and transform the predictor variables into a set of uncorrelated variables. We analyze the principal component scores of temperature and rainfall in detail. We find that the first two to three principal component scores can explain more than three-fourths of the variation in the weather data.

We use smoothing parameter values and t-test criteria to find the final semiparametric regression model. The selected model is compared with a set of other yield

forecasting models. We find that the selected semiparametric regression model outperforms the parametric linear regression models in terms of mean squared error. Even when we use forecasted weather data rather than observed weather data, the selected semiparametric model still provides the best predictions. We can conclude that allowing for the time-varying relationships of the data both within and between years together with the functional principal component technique can improve the forecasting performance over the ordinary parametric regression approach. In addition, we estimate the pointwise confidence interval, which provides the range of possible values of the yield within a year. With this estimated confidence interval, we can investigate the effect of the yield uncertainty in the decision planning model.

In Chapter 3, we develop a price forecasting model under a futures-based framework. This approach is based on a relationship that cash price equals futures price plus commodity basis. Commodity basis is the difference between local cash price and futures contract price, and it reflects the local market conditions. We focus on forecasting commodity basis rather than cash price because commodity basis does not change dramatically from year to year. We analyze the basis of corn and soybean. We find that both corn and soybean basis fluctuate throughout the year with five local maxima and one local minimum. We apply the functional model-based clustering approach to the standardized commodity basis to estimate the commodity basis distribution over one year. This technique allows us to determine the confidence interval of the commodity basis at any time point in the forecasted year. Therefore, our price forecasting model is distinct from other futures-based models that usually calculate the expected commodity basis. We investigate the effects from the number of clusters and the dimension of the spline basis. We find that only the dimension of the spline basis affects the prediction. Because the commodity basis forecasts are in standardized scale, we propose a calibration procedure that rescales the predicted commodity basis back to their normal scale. We obtain the forecasted cash price by

adding the futures price to the forecasted commodity basis.

In Chapter 4, we develop stochastic and heuristic crop decision planning models. Our models consider the complete planning process including crop selection, acreage allocation, planting and harvesting scheduling, storing, and selling. Consequently, our research is different from other studies that consider portions of the process. Since the decisions in each part impact other parts and decisions on earlier parts are irreversible, making decision on each part individually cannot deliver the best overall results. Our models are based on the entire process and this method can significantly enhance the overall results. The objective of the models is to maximize the expected profit. The planning horizon covers the duration from the beginning of the planting periods through the end of year. Yield and price are the uncertainties in our study. We assume that yield is observed at the beginning and the middle of the harvest season. Price, on the other hand, is observed at the middle and the end of the harvest season, and the end of year. These assumptions result in a five-stage decision problem. We utilize the forecasted yield and price from Chapters 2 and 3, respectively. Even though we forecast confidence intervals for yield and price, we assume discrete outcomes. We discretize the intervals to avoid having an infinite number of scenarios. The main objective of our planning models concentrates on coordinating decisions between stages in the process. Incorporating the predicted confidence intervals into the models would result in a much longer computational time. In addition, solving the problem with an infinite number of scenarios is not the main objective of this research.

The first decision planning model is the stochastic linear programming model (SLP). This model determines the optimal solution and provides the policy or strategy for every possible scenario. However, the computational effort increases dramatically as the number of stages increases. Therefore, we develop simple heuristic models, each of which delivers a feasible solution in polynomial time. We propose three heuristic

models, namely Greedy Deterministic Approach (GDA), Greedy Optimistic Deterministic Approach (GOA), and Greedy Pessimistic Deterministic Approach (GPA). These heuristic models are based on greedy algorithms but use different criteria to determine a representative of future uncertainty. GDA uses the most likely outcome, while GOA chooses the best outcome. On the other hand, GPA selects the worst outcome.

Small and medium examples are examined in the numerical study. In the small example, we assume that both corn and soybean have the same uncertainty outcome in every stage. In contrast, in the medium example, we assume outcomes of corn and soybean are independent. We find that both examples result in the same acreage allocation decision. SLP gives the highest average profit. GPA gives the return very close to SLP. GOA provides the smallest average profit. GOA generates a moderate return. However, in terms of average computational time, the heuristic models can solve the problem much faster than the stochastic programming model when the problem size is large. We also investigate the robustness of the model results to the uncertainties and prior probabilities. We find that our models are robust to the prices but not to the yield. In addition, SLP and GPA are robust to the prior probabilities but GDA is not.

5.2 Future Research

In yield forecasting, we use only GDP and monthly data of temperature and rainfall as predictors in our model. There are many other factors that can affect crop yield such as soil properties, humidity, cultivation techniques (tillage and non-tillage), and pesticide. One possible extension is to use more predictors and finer time grid (e.g. daily temperature and rainfall) in the regression model. Even though adding more variables will increase the dimensionality of the model, we can use a variable selection technique to select only the important variables. This extension may enhance the

accuracy of the predicted yield.

In this dissertation, we apply FPCA to temperature and rainfall data separately; hence, multicollinearity between the principal component scores of temperature and rainfall still exists. Therefore, another challenging extension would be to apply FPCA to bivariate or multivariate functional data. Applying FPCA to all predictor variables together will result in a set of uncorrelated variables. Consequently, we can precisely estimate the contribution of each principal component on the forecasted yield.

In price forecasting, the accuracy of forecasted cash price also depends on futures price. Since futures price changes continuously, we cannot get the exact predicted cash price. This may affect the decision planning. An estimation of the range of the possible futures price values would be useful.

In crop decision planning model, there are several potential directions for future research. A natural extension would be to add more stages to the problem in order to account for the other decisions that farmers make. Considering more crop types that farmers can choose is another way of expanding this research.

One could also consider long-term decision planning rather than short-term. The capacity expansion, i.e. land, labor, equipment, would be a crucial decision in this case.

Incorporating Bayesian analysis in the future work could improve the accuracy of the probabilities used in the planning model since this technique updates the prior probabilities with the information we receive from the data (see Carlin and Louis 2000, Gelman et al. 2004).

It is also possible to utilize the functional values of the forecasted yield and price rather than using only the discrete values since using only some selected values discards the possible scenarios that may occur and may lead to incorrect decisions. One possible technique to handle this extension in stochastic programming is Sample Average Approximation (see Ahmed and Shapiro 2002).

APPENDIX A

DETAILED GREEDY ALGORITHMS

Step 0: Initialization

Step 0.1: Initialize the solution set that will hold the model solution as an empty set.

Step 0.2: Initialize the realization set that will hold the realization of the uncertainties as an empty set.

Step 1: Solving the first stage

Step 1.1: Determine the outcomes of future uncertainties using the pre-specified criterion.

Step 1.2: Solve the deterministic model using the uncertainty outcomes from Step 1.1.

Step 1.3: Add the solution of the first-stage variables from Step 1.2 to the solution set.

Step 2: Solving the second stage

Step 2.1: Observe the realization of yield1 and add this outcome to the realization set.

Step 2.2: Determine the outcomes of yield2, price_h1, price_h2, and price_e using the pre-specified criterion.

Step 2.3: Set the values of variables in solution set as parameters.

Step 2.4: Re-solve the deterministic model using parameters from Step 2.3 and realization of uncertainty in the realization set from Step 2.1 and the remaining uncertainty outcomes from Step 2.2.

Step 2.5: Add the solution of the second-stage variables from Step 2.4 to the solution set.

Step 3: Solving the third stage

Step 3.1: Observe the realizations of yield2 and price_h1 and add these outcomes to the realization set.

Step 3.2: Determine the outcomes of price_h2, and price_e using the pre-specified criterion.

Step 3.3: Set the values of variables in solution set as parameters.

Step 3.4: Re-solve the deterministic model using parameters from Step 3.3 and realizations of uncertainties in the realization set from Step 3.1 and the remaining uncertainty outcomes from Step 3.2.

Step 3.5: Add the solution of the third-stage variables from Step 3.4 to the solution set.

Step 4: Solving the fourth stage

Step 4.1: Observe the realization of price_h2 and add this outcome to the realization set.

Step 4.2: Determine the outcome of price_e using the pre-specified criterion.

Step 4.3: Set the values of variables in solution set as parameters.

Step 4.4: Re-solve the deterministic model using parameters from Step 4.3 and realizations of uncertainties in the realization set from Step 4.1 and the remaining uncertainty outcomes from Step 4.2.

Step 4.5: Add the solution of the fourth-stage variables from Step 4.4 to the solution set.

Step 5: Solving the fifth stage

Step 5.1: Observe the realization of price_e and add this outcome in the realization set.

Step 5.2: Set the values of variables in solution set as parameters.

Step 5.3: Re-solve the deterministic model using parameters from Step 5.2 and realizations of uncertainties in the realization set from Step 5.1.

Step 5.4: Add the solution of the fifth-stage variables form Step 5.3 to the solution set.

APPENDIX B

PRIOR PROBABILITIES

In Section 4.4.2, the conditional probabilities are defined as:

$p_1(s_1)$	=	Probability that outcome of s_1 will occur
$p_2(s_2, s_3\|s_1)$	=	Conditional probability that outcomes s_2 and s_3 will occur given outcome s_1
$p_3(s_4\|s_1, s_2, s_3)$	=	Conditional probability that outcome s_4 will occur given outcomes s_1, s_2, s_3
$p_4(s_5\|s_1, s_2, s_3, s_4)$	=	Conditional probability that outcome s_5 will occur given outcomes s_1, s_2, s_3, s_4.

In addition, we define

$p(s_2\|s_1)$	=	Probability that outcome s_2 will occur given outcome s_1
$p(s_3\|s_1)$	=	Probability that outcome s_3 will occur given outcome s_1.

In this research, we assume that

1. Outcome of yield2 (S_2) does not depend on any outcome of previous uncertainties.

2. Yield2 (S_2) and price_h1 (S_3) are conditional independent given yield1 (S_1)

3. Outcome of price_h1 (S_3) depends only on the outcome of yield1 (S_1).

4. Outcome of price_h2 (S_4) depends only on the outcome of yield2 (S_2).

5. Outcome of price_e (S_5) depends only on the outcome of price_h2 (S_4).

Assumption 1 implies that

$$p(s_2|s_1) = p(s_2).$$

Assumption 2 implies that

$$p_2(s_2, s_3|s_1) = p(s_2|s_1) \times p(s_3|s_1).$$

Assumption 4 implies that

$$p_3(s_4|s_1, s_2, s_3) = p_3(s_4|s_2).$$

Assumption 5 implies that

$$p_4(s_5|s_1, s_2, s_3, s_4) = p_4(s_5|s_4).$$

The prior probabilities used in the crop decision planning model are:

Table 22: Prior probabilities

Outcome	Probability
$p_1(s_1 = High)$	0.20
$p_1(s_1 = Medium)$	0.60
$p_1(s_1 = Low)$	0.20
$p(s_2 = High)$	0.50
$p(s_2 = Medium)$	0.30
$p(s_2 = Low)$	0.20
$p(s_3 = High \mid s_1 = High)$	0.15
$p(s_3 = Medium \mid s_1 = High)$	0.30
$p(s_3 = Low \mid s_1 = High)$	0.55
$p(s_3 = High \mid s_1 = Medium)$	0.20
$p(s_3 = Medium \mid s_1 = Medium)$	0.60
$p(s_3 = Low \mid s_1 = Medium)$	0.20
$p(s_3 = High \mid s_1 = Low)$	0.55
$p(s_3 = Medium \mid s_1 = Low)$	0.30
$p(s_3 = Low \mid s_1 = Low)$	0.15
$p_3(s_4 = High \mid s_2 = High)$	0.20
$p_3(s_4 = Medium \mid s_2 = High)$	0.30
$p_3(s_4 = Low \mid s_2 = High)$	0.50
$p_3(s_4 = High \mid s_2 = Medium)$	0.25
$p_3(s_4 = Medium \mid s_2 = Medium)$	0.50
$p_3(s_4 = Low \mid s_2 = Medium)$	0.25
$p_3(s_4 = High \mid s_2 = Low)$	0.50
$p_3(s_4 = Medium \mid s_2 = Low)$	0.30
$p_3(s_4 = Low \mid s_2 = Low)$	0.20
$p_4(s_5 = High \mid s_4 = High)$	0.70
$p_4(s_5 = Medium \mid s_4 = High)$	0.20
$p_4(s_5 = Low \mid s_4 = High)$	0.10
$p_4(s_5 = High \mid s_4 = Medium)$	0.20
$p_4(s_5 = Medium \mid s_4 = Medium)$	0.60
$p_4(s_5 = Low \mid s_4 = Medium)$	0.20
$p_4(s_5 = High \mid s_4 = Low)$	0.10
$p_4(s_5 = Medium \mid s_4 = Low)$	0.20
$p_4(s_5 = Low \mid s_4 = Low)$	0.70

Table 23: Adjusted prior probabilities - equal chance for yield1 and yield2 (for sensitivity analysis)

Outcome	Probability
$p_1(s_1 = High)$	1/3
$p_1(s_1 = Medium)$	1/3
$p_1(s_1 = Low)$	1/3
$p(s_2 = High)$	1/3
$p(s_2 = Medium)$	1/3
$p(s_2 = Low)$	1/3
$p(s_3 = High\|s_1 = High)$	0.15
$p(s_3 = Medium\|s_1 = High)$	0.30
$p(s_3 = Low\|s_1 = High)$	0.55
$p(s_3 = High\|s_1 = Medium)$	0.20
$p(s_3 = Medium\|s_1 = Medium)$	0.60
$p(s_3 = Low\|s_1 = Medium)$	0.20
$p(s_3 = High\|s_1 = Low)$	0.55
$p(s_3 = Medium\|s_1 = Low)$	0.30
$p(s_3 = Low\|s_1 = Low)$	0.15
$p_3(s_4 = High\|s_2 = High)$	0.20
$p_3(s_4 = Medium\|s_2 = High)$	0.30
$p_3(s_4 = Low\|s_2 = High)$	0.50
$p_3(s_4 = High\|s_2 = Medium)$	0.25
$p_3(s_4 = Medium\|s_2 = Medium)$	0.50
$p_3(s_4 = Low\|s_2 = Medium)$	0.25
$p_3(s_4 = High\|s_2 = Low)$	0.50
$p_3(s_4 = Medium\|s_2 = Low)$	0.30
$p_3(s_4 = Low\|s_2 = Low)$	0.20
$p_4(s_5 = High\|s_4 = High)$	0.70
$p_4(s_5 = Medium\|s_4 = High)$	0.20
$p_4(s_5 = Low\|s_4 = High)$	0.10
$p_4(s_5 = High\|s_4 = Medium)$	0.20
$p_4(s_5 = Medium\|s_4 = Medium)$	0.60
$p_4(s_5 = Low\|s_4 = Medium)$	0.20
$p_4(s_5 = High\|s_4 = Low)$	0.10
$p_4(s_5 = Medium\|s_4 = Low)$	0.20
$p_4(s_5 = Low\|s_4 = Low)$	0.70

Table 24: Adjusted prior probabilities - equal chance for every uncertainty (for sensitivity analysis)

Outcome	Probability
$p_1(s_1 = High)$	1/3
$p_1(s_1 = Medium)$	1/3
$p_1(s_1 = Low)$	1/3
$p(s_2 = High)$	1/3
$p(s_2 = Medium)$	1/3
$p(s_2 = Low)$	1/3
$p(s_3 = High\|s_1 = High)$	1/3
$p(s_3 = Medium\|s_1 = High)$	1/3
$p(s_3 = Low\|s_1 = High)$	1/3
$p(s_3 = High\|s_1 = Medium)$	1/3
$p(s_3 = Medium\|s_1 = Medium)$	1/3
$p(s_3 = Low\|s_1 = Medium)$	1/3
$p(s_3 = High\|s_1 = Low)$	1/3
$p(s_3 = Medium\|s_1 = Low)$	1/3
$p(s_3 = Low\|s_1 = Low)$	1/3
$p_3(s_4 = High\|s_2 = High)$	1/3
$p_3(s_4 = Medium\|s_2 = High)$	1/3
$p_3(s_4 = Low\|s_2 = High)$	1/3
$p_3(s_4 = High\|s_2 = Medium)$	1/3
$p_3(s_4 = Medium\|s_2 = Medium)$	1/3
$p_3(s_4 = Low\|s_2 = Medium)$	1/3
$p_3(s_4 = High\|s_2 = Low)$	1/3
$p_3(s_4 = Medium\|s_2 = Low)$	1/3
$p_3(s_4 = Low\|s_2 = Low)$	1/3
$p_4(s_5 = High\|s_4 = High)$	1/3
$p_4(s_5 = Medium\|s_4 = High)$	1/3
$p_4(s_5 = Low\|s_4 = High)$	1/3
$p_4(s_5 = High\|s_4 = Medium)$	1/3
$p_4(s_5 = Medium\|s_4 = Medium)$	1/3
$p_4(s_5 = Low\|s_4 = Medium)$	1/3
$p_4(s_5 = High\|s_4 = Low)$	1/3
$p_4(s_5 = Medium\|s_4 = Low)$	1/3
$p_4(s_5 = Low\|s_4 = Low)$	1/3

APPENDIX C

ANOVA TABLES

Tables 25 to 28 display the ANOVAs for the small example while Tables 29 to 32 display for the medium example.

Table 25: ANOVA for SLP of the small example

Source	**DF**	**SS**	**MS**	**F**	**P**
Yield1	1	3879868133	3879868133	701.43	0.024†
Price_h1	1	266980649	266980649	48.27	0.091‡
Price_h2	1	242443195	242443195	43.83	0.095‡
Price_e	1	283913148	283913148	51.33	0.088‡
Yield1*Price_h1	1	5660414	5660414	1.02	0.496
Yield1*Price_h2	1	11640089	11640089	2.10	0.384
Yield1*Price_e	1	6333741	6333741	1.15	0.478
Price_h1*Price_h2	1	257894934	257894934	46.62	0.093‡
Price_h1*Price_e	1	257559568	257559568	46.56	0.093‡
Price_h2*Price_e	1	272788083	272788083	49.32	0.090‡
Yield1*Price_h1*Price_h2	1	5318639	5318639	0.96	0.506
Yield1*Price_h1*Price_e	1	5177411	5177411	0.94	0.511
Yield1*Price_h2*Price_e	1	5353659	5353659	0.97	0.505
Price_h1*Price_h2*Price_e	1	251261413	251261413	45.42	0.094‡
Error	1	5531375	5531375		
Total	15	5757724452			

† significance at $\alpha = 0.05$, ‡ significance at $\alpha = 0.10$

Table 26: ANOVA for GDA of the small example

Source	DF	SS	MS	F	P
Yield1	1	3533279440	3533279440	772.04	0.023†
Price_h1	1	448362626	448362626	97.97	0.064‡
Price_h2	1	159528141	159528141	34.86	0.107
Price_e	1	220034354	220034354	48.08	0.091‡
Yield1*Price_h1	1	8082080	8082080	1.77	0.411
Yield1*Price_h2	1	10601080	10601080	2.32	0.370
Yield1*Price_e	1	4206847	4206847	0.92	0.513
Price_h1*Price_h2	1	262947954	262947954	57.46	0.084‡
Price_h1*Price_e	1	230354989	230354989	50.33	0.089‡
Price_h2*Price_e	1	207304708	207304708	45.30	0.094‡
Yield1*Price_h1*Price_h2	1	4987607	4987607	1.09	0.486
Yield1*Price_h1*Price_e	1	5916448	5916448	1.29	0.459
Yield1*Price_h2*Price_e	1	4517006	4517006	0.99	0.502
Price_h1*Price_h2*Price_e	1	230442874	230442874	50.35	0.089‡
Error	1	4576562	4576562		
Total	15	5335142715			

† significance at $\alpha = 0.05$, ‡ significance at $\alpha = 0.10$

Table 27: ANOVA for GOA of the small example

Source	DF	SS	MS	F	P
Yield1	1	2916071820	2916071820	337.63	0.035†
Price_h1	1	296212153	296212153	34.30	0.108
Price_h2	1	157748702	157748702	18.26	0.146
Price_e	1	158327737	158327737	18.33	0.146
Yield1*Price_h1	1	8838907	8838907	1.02	0.496
Yield1*Price_h2	1	11436436	11436436	1.32	0.455
Yield1*Price_e	1	387307	387307	0.04	0.867
Price_h1*Price_h2	1	300613873	300613873	34.81	0.107
Price_h1*Price_e	1	300965942	300965942	34.85	0.107
Price_h2*Price_e	1	155240635	155240635	17.97	0.147
Yield1*Price_h1*Price_h2	1	9034984	9034984	1.05	0.493
Yield1*Price_h1*Price_e	1	8670699	8670699	1.00	0.499
Yield1*Price_h2*Price_e	1	1264894	1264894	0.15	0.767
Price_h1*Price_h2*Price_e	1	297633747	297633747	34.46	0.107
Error	1	8636957	8636957		
Total	15	4631084792			

† significance at $\alpha = 0.05$

Table 28: ANOVA for GPA of the small example

Source	**DF**	**SS**	**MS**	**F**	**P**
Yield1	1	4458538044	4458538044	4499.52	0.009†
Price_h1	1	291697536	291697536	294.38	0.037†
Price_h2	1	373691041	373691041	377.13	0.033†
Price_e	1	317594991	317594991	320.51	0.036†
Yield1*Price_h1	1	1409634	1409634	1.42	0.444
Yield1*Price_h2	1	27697	27697	0.03	0.895
Yield1*Price_e	1	1424538	1424538	1.44	0.443
Price_h1*Price_h2	1	296988180	296988180	299.72	0.037†
Price_h1*Price_e	1	295891258	295891258	298.61	0.037†
Price_h2*Price_e	1	305407955	305407955	308.22	0.036†
Yield1*Price_h1*Price_h2	1	1230081	1230081	1.24	0.466
Yield1*Price_h1*Price_e	1	1292485	1292485	1.30	0.458
Yield1*Price_h2*Price_e	1	1802870	1802870	1.82	0.406
Price_h1*Price_h2*Price_e	1	288901748	288901748	291.56	0.037†
Error	1	990891	990891		
Total	15	6636888948			

† significance at $\alpha = 0.05$

Table 29: ANOVA for SLP of the medium example

Source	**DF**	**SS**	**MS**	**F**	**P**
Yield1	1	3885067545	3885067545	740.65	0.023†
Price_h1	1	257246313	257246313	49.04	0.090‡
Price_h2	1	247885659	247885659	47.26	0.092‡
Price_e	1	285940998	285940998	54.51	0.086‡
Yield1*Price_h1	1	5478713	5478713	1.04	0.493
Yield1*Price_h2	1	12362959	12362959	2.36	0.368
Yield1*Price_e	1	5788619	5788619	1.10	0.484
Price_h1*Price_h2	1	248375394	248375394	47.35	0.092‡
Price_h1*Price_e	1	250540621	250540621	47.76	0.091‡
Price_h2*Price_e	1	271101823	271101823	51.68	0.088‡
Yield1*Price_h1*Price_h2	1	5410578	5410578	1.03	0.495
Yield1*Price_h1*Price_e	1	5232199	5232199	1.00	0.500
Yield1*Price_h2*Price_e	1	6467180	6467180	1.23	0.467
Price_h1*Price_h2*Price_e	1	251882451	251882451	48.02	0.091‡
Error	1	5245474	5245474		
Total	15	5744026526			

† significance at $\alpha = 0.05$, ‡ significance at $\alpha = 0.10$

Table 30: ANOVA for GDA of the medium example

Source	DF	SS	MS	F	P
Yield1	1	3463075923	3463075923	581.04	0.035†
Price_h1	1	273937918	273937918	45.96	0.093‡
Price_h2	1	160988262	160988262	27.01	0.121
Price_e	1	188927362	188927362	31.70	0.112
Yield1*Price_h1	1	4559463	4559463	0.76	0.543
Yield1*Price_h2	1	10069992	10069992	1.69	0.417
Yield1*Price_e	1	4131971	4131971	0.69	0.558
Price_h1*Price_h2	1	272161237	272161237	45.66	0.094‡
Price_h1*Price_e	1	276665338	276665338	46.42	0.093‡
Price_h2*Price_e	1	179104823	179104823	30.05	0.115
Yield1*Price_h1*Price_h2	1	5788451	5788451	0.97	0.505
Yield1*Price_h1*Price_e	1	5587456	5587456	0.94	0.510
Yield1*Price_h2*Price_e	1	3478579	3478579	0.58	0.585
Price_h1*Price_h2*Price_e	1	268123595	268123595	44.99	0.094‡
Error	1	5960092	5960092		
Total	15	5122560463			

† significance at $\alpha = 0.05$, ‡ significance at $\alpha = 0.10$

Table 31: ANOVA for GOA of the medium example

Source	DF	SS	MS	F	P
Yield1	1	2956356794	2956356794	334.42	0.035†
Price_h1	1	298586304	298586304	33.78	0.108
Price_h2	1	168409926	168409926	19.05	0.143
Price_e	1	153433186	153433186	17.36	0.150
Yield1*Price_h1	1	8344299	8344299	0.94	0.509
Yield1*Price_h2	1	12472163	12472163	1.41	0.445
Yield1*Price_e	1	601299	601299	0.07	0.838
Price_h1*Price_h2	1	297630986	297630986	33.67	0.109
Price_h1*Price_e	1	301282633	301282633	34.08	0.108
Price_h2*Price_e	1	154899431	154899431	17.52	0.149
Yield1*Price_h1*Price_h2	1	9087060	9087060	1.03	0.496
Yield1*Price_h1*Price_e	1	9805885	9805885	1.11	0.484
Yield1*Price_h2*Price_e	1	1062198	1062198	0.12	0.788
Price_h1*Price_h2*Price_e	1	304864870	304864870	34.49	0.107
Error	1	8840216	8840216		
Total	15	4685677250			

† significance at $\alpha = 0.05$

Table 32: ANOVA for GPA of the medium example

Source	**DF**	**SS**	**MS**	**F**	**P**
Yield1	1	4463059006	4463059006	3020.81	0.012†
Price_h1	1	281551117	281551117	190.57	0.046†
Price_h2	1	386309591	386309591	261.47	0.039†
Price_e	1	320135115	320135115	216.68	0.043†
Yield1*Price_h1	1	1146805	1146805	0.78	0.540
Yield1*Price_h2	1	338090	338090	0.23	0.716
Yield1*Price_e	1	1349651	1349651	0.91	0.514
Price_h1*Price_h2	1	294965855	294965855	199.65	0.045†
Price_h1*Price_e	1	298824811	298824811	202.26	0.045†
Price_h2*Price_e	1	309863673	309863673	209.73	0.044†
Yield1*Price_h1*Price_h2	1	1280711	1280711	0.87	0.523
Yield1*Price_h1*Price_e	1	1429471	1429471	0.97	0.505
Yield1*Price_h2*Price_e	1	2186968	2186968	1.48	0.438
Price_h1*Price_h2*Price_e	1	294374144	294374144	199.25	0.045†
Error	1	1477440	1477440		
Total	15	6658292448			

† significance at $\alpha = 0.05$

REFERENCES

[1] Adam, B.D., Garcia, P., Hauser, R.J. (1996), "The value of information to hedgers in the presence of futures and options," *Review of Agricultural Economics*, 18(3), 437-447.

[2] Ahmed, S., Shapiro, A. (2002), "The sample average approximation method for stochastic programs with integer recourse," ISyE Technical Report, Georgia Institute of Technology, Altanta.

[3] Antonovitz, F., Roe, T. (1986), "A theory and empirical approach to the value of information in risky markets," *Review of Economics & Statistics*, 86(1), 105-114.

[4] Babcock, B.A. (1990), "Acreage decisions under marketing quotas and yield uncertainty," *American Journal of Agricultural Economics*, 72(4), 958-965.

[5] Baier, W. (1979), "Note on the terminology of crop-weather models," *Agricultural Meteorology*, 20(2), 137-145.

[6] Ballal, M.E., Siddig, E.A.E., Elfadl, M.A., Luukkanen, O. (2005), "Relationship between environmental factors, tapping dates, tapping intensity and gum arabic yield of an Acacia Senegal plantation in Western Sudan," *Journal of Arid Environments*, 63(2), 379-389.

[7] Banfield, J.D., Raftery, A.E. (1993), "Model-based gaussian and non-gaussian clustering," *Biometrics*, 49, 803-821.

[8] Bannayan, M., Crout, N.M.J. (1999), "A stochastic modelling approach for real-time forecasting of winter wheat yield," *Field Crops Research*, 62, 85-95.

[9] Batts, G.R., Morison, J.I.L., Ellis, R.H., Hadley, P., Wheeler, T.R. (1997), "Effects of CO_2 and temperature on growth and yield of crops of winter wheat over four seasons," *European Journal of Agronomy*, 7, 43-52.

[10] Boken, V.K. (2000), "Forecasting spring wheat yield using time series analysis: A case study for the Canadian prairies," *Agronomy Journal*, 92(6), 1047-1053.

[11] Bureau of Economic Analysis, United State Department of Commerce (2006), "Gross domestic product (GDP) by industry data," http://www.bea.gov/industry/gdpbyind_data.htm, last accessed on May 19, 2007.

[12] Bureau of Labor Statistics, United State Department of Labor (2005), "Career guide to industires (CGI), 2006-07 edition," http://www.bls.gov/oco/cg/cgs001.htm, last accessed on May 19, 2007.

[13] Butterworth, K. (1985), "Practical application of linear/integer programming in agriculture," *The Journal of the Operational Research Society*, 36(2), 99-107.

[14] Byerlee, D., Anderson, J.R. (1982), "Risk, utility, and the value of information in farmer decision making," *Review of Marketing and Agricultural Economics*, 50, 231-246.

[15] Carlin B.P., Louis T.A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall/CRC, Boca Raton.

[16] Celeux, G., Govaert, G. (1995), "Gaussian parsimonious clustering models," *The Journal of the Pattern Recognition Society*, 28, 781-793.

[17] Chavas, J.P., Holt, M.T. (1990), "Acreage decisions under risk: The case of corn and soybeans," *American Journal of Agricultural Economics*, 72(3), 529-538.

[18] Chicago Board of Trade (2000), "Understanding basis," http://www.cbot.com/cbot/docs/46577.pdf, last accessed on May 17, 2007.

[19] Cocks, K.D. (1968), "Discrete stochastic programming," *Management Science*, 15(1), 72-79.

[20] Commodity Credit Corporation, United State Department of Agriculture, http://www.fsa.usda.gov/FSA/webapp?area=about&subject=landing&topic=sao-cc-ac, last accessed on May 13, 2007.

[21] Dasgupta, A., Raftery, A.E. (1998), "Detecting features in spatial point porcesses with clutter via model-based clustering," *Journal of the American Statistical Association*, 93, 294-302.

[22] De la Rosa, D., Cardona, F., Almorza, J. (1981), "Crop yield prediction based on properties of soils in Sellilla, Spain," *Geoderma*, 25(3-4), 267-274.

[23] De la Torre, F., Black, M.J. (2001), "Robust principal component analysis for computer vision," *International Conference on Computer Vision, ICCV-2001*, 1, 362-369.

[24] Dow, J.C.R. (1940), "A theoretical account of futures markets," *The Review of Economic Studies*, 7, 185-195.

[25] Eales, J.S., Engel, B.K., Hauser, R.J., Thompson, S.R. (1990), "Grain price expectations of Illinois farmers and grain merchandisers," *American Journal of Agricultural Economics*, 72(3), 701-708.

[26] Economic History Services, http://eh.net/hmit/gdp/, last accessed on May 17, 2007.

[27] Economic Research Service, United State Department of Agriculture (2006), "Agricultural resources and environment indicators, 2006 edition," http://www.ers.usda.gov/publications/arei/eib16/eib16_1-1.pdf, last accessed on May 19, 2007.

[28] Ethridge, M.D., White, F.C., Kannan, D. (1975), "Optimizing seed acreage: Decision making with production and utilization uncertainties," *American Journal of Agricultural Economics*, 57(3), 439-449.

[29] Fackler, P.L., Livingston, M.J. (2002), "Optimal storage by crop producers," *American Journal of Agricultural Economics*, 84(3), 645-659.

[30] Fokkens, B., Puylaert, M. (1981), "A linear programming model for daily harvesting operations at the large-scale farm of the IJsselmeerpolders Development Authority," *The Journal of the Operational Research Society*, 32(7), 535-547.

[31] Freckleton, R.P., Watkinson, A.R., Webb, D.J., Thomas, T.H. (1999), "Yield of sugar beet in relation to weather and nutrients," *Agricultural and Forest Meteorology*, 93(1), 39-51.

[32] Garcia-Paredes, J.D., Olson, K.R., Lang, J.M. (2000), "Predicting corn and soybean productivity for Illinois soils," *Agricultural Systems*, 64(3), 151-170.

[33] Gardner, B.L. (1976), "Futures prices in supply analysis," *American Journal of Agricultural Economics*, 58, 81-84.

[34] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton.

[35] Glen, J.J. (1987), "Mathematical models in farm planning: A survey," *Operations Research*, 35(5), 641-666.

[36] Greenwald, R., Bergin, M.H., Xu, J., Cohan, D., Hoogenboom, G., Chameides, W.L. (2006), "The influence of aerosols on crop production: A study using the CERES crop model," *Agricultural Systems*, 89(2-3), 390-413.

[37] Guise, J.W.B. (1969), "Factors associated with variation in the aggregate average yield of New Zealand wheat (1918-1967)," *American Journal of Agricultural Economics*, 51, 866-881.

[38] Hauser, R.J., Garcia, P., Tumblin, A.D. (1990), "Basis expectations and soybean hedging effectiveness," *North Central Journal of Agricultural Economics*, 12(1), 125-136.

[39] Heady, E.O. (1954), "Simplified presentation and logical aspects of linear programming technique," *Journal of Farm Economics*, 36(5), 1035-1048.

[40] Hoffman, L.A. (2005), "Forecasting the counter-cyclical payment rate for U.S. corn: An application of the futures price forecasting model," Outlook Report No. FDS-05a-01, Economic Research Service, http://www.ers.usda.gov/publications/FDS/JAN05/fds05a01/fds05a01.pdf, last accessed on May 17, 2007.

[41] Hoogenboom, G. (2000), "Contribution of agrometeorology to the simulation of crop production and its applications," *Agricultural and Forest Meteorology*, 103, 137-157.

[42] Horie, T., Yajima, M., Nakagawa, H. (1992), "Yield forecasting," *Agricultural Systems*, 40(1-3), 211-236.

[43] House, C.C. (1979), "Forecasting corn yields: A comparison study using 1977 Missouri data," Unnumbered report, Statistical Research Division, United States Department of Agriculture.

[44] Huda, A.K.S., Ghildyal, B.P., Tomar, V.S. (1976), "Contribution of climatic variables in predicting maize yield under monsoon conditions," *Agricultural Meteorology*, 17(1), 33-47.

[45] Irwin, S.H., Gerlow, M.E., Liu, T.R. (1994), "The forecasting performance of livestock futures prices: A comparison to USDA expert predictions," *Journal of Futures Markets*, 14(7), 861-875.

[46] Isengildina, O., Irwin, S.H., Good, D.L. (2004), "Evaluation of USDA interval forecasts of corn and soybean prices," *American Journal of Agricultural Economics*, 86(4), 990-1004.

[47] Itoh, T., Ishii, H., Nanseki, T. (2003), "A model of crop planning under uncertainty in agricultural management," *International Journal of Production Economics*, 81-82, 555-558.

[48] James, G.M., Sugar, C.A. (2003), "Clustering for sparsely sampled functional data," *Journal of the American Statistical Association*, 98, 397-408.

[49] Jiang, D., Yang, X., Clinton, N., Wang, N. (2004), "An artificial neural network model for estimating crop yields using remotely sensed information," *International Journal of Remote Sensing*, 25(9), 1723-1732.

[50] Jiao, Z., Higgins, A.J., Prestwidge, D.B. (2005), "An integrated statistical and optimisation approach to increasing sugar production within a mill region," *Computers and Electronics in Agriculture*, 48(2), 170-181.

[51] Jones, P.C., Lowe, T.J., Traub, R.D. (2002), "Matching supply and demand: The value of a second chance in producing seed corn," *Review of Agricultural Economics*, 24(1), 222-238.

[52] Just, R.E., Rausser, G.C. (1981), "Commodity price forecasting with large-scale econometric modes and the futures market," *American Journal of Agricultural Economics*, 63(2), 197-208.

[53] Kandiannan, K., Chandaragiri, K.K., Sankaran, N., Balasubramanian, T.N., Kailasam, C. (2002), "Crop-weather model for turmeric yield forecasting for Coimbatore District, Tamil Nadu, India," *Agricultural and Forest Meteorology*, 112, 133-137.

[54] Kaspar, T.C., Colvin, T.S., Jaynes, D.B., Karlen, D.L., James, D.E., Meek, D.W., Pulido, D., Butler, H. (2003), "Relationship between six years of corn yields and terrain attributes," *Precision Agriculture*, 4(1), 87-101.

[55] Kass, R.E., Raftery, A.E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773-795.

[56] Kastens, T.L., Jones, R., Schroeder, T.C. (1998), "Futures-based price forecasts for agricultural producers and businesses," *Journal of Agricultural and Resource Economics*, 23(1), 294-307.

[57] Kaul, M., Hill, R.L., Walthall, C. (2005), "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, 85(1), 1-18.

[58] Kazaz, B. (2004), "Production planning under yield and demand uncertainty with yield-dependent cost and price," *Manufacturing & Service Operations Management*, 6(3), 209-224.

[59] Kenyon, D., Jones, E., McGuirk, A. (1993), "Forecasting performance of corn and soybean harvest futures contracts," *American Journal of Agricultural Economics*, 75, 399-407.

[60] Kenyon, D., Kingsley, S.E. (1973), "An Analysis of Anticipatory Short Hedging Using Predicted Harvest Basis," *Southern Journal of Agricultural Economics*, 5, 199-203.

[61] Kenyon, D., Lucas, K. (1998), "Soybean pricing guide," Department of Agricultural and Applied Economics, Virginia Tech, Virginia, USA, REAP Report No. 37., http://www.reap.vt.edu/publications/reports/r37.PDF, last accessed on May 17, 2007.

[62] Krog, D.R. (1988), Plant-process model corn yield forecasts for Iowa, Ph.D. Dissertation, Iowa State University, Iowa.

[63] Lai, J., Myers, R.J., Hanson, S.D. (2003), "Optimal on-farm grain storage by risk-averse farmers," *Journal of Agricultural and Resource Economics*, 28(3), 558-579.

[64] Lee, R. (1999), Modeling corn yields in Iowa using time series analysis of AVHRR data and vegetation phenological metrics, Ph.D. Dissertation, University of Kansas, Kansas.

[65] Lence, S.H., Hayes, D.J. (2002), "U.S. farm policy and the volatility of commodity prices and farm revenues," *American Journal of Agricultural Economics*, 84(2), 335-351.

[66] Liew, V. K.-S., Shitan, M., Hussian, H. (2003), "Time series modelling and forecasting of Sarawak black pepper price," *Jurnal Akademik*, 39-55.

[67] Lomas, J., Herrera, H. (1985), "Weather and rice yield relationships in tropical Costa Rica," *Agricultural and Forest Meteorology*, 35(1-4), 133-151.

[68] Lowe, T.J., Preckel, P.V. (2004), "Decision technologies for agribusiness problems: A brief review of selected literature and a call for research," *Manufacturing & Service Operations Management*, 6(3), 201-208.

[69] Maatman, A., Schweigman, C., Ruijs, A., van der Vlerk, M.H. (2002), "Modeling farmers' response to uncertain rainfall in Burkina Faso: A stochastic programming approach," *Operations Research*, 50(3), 399-414.

[70] Mantis, J.H., Birkett, T., Boudreaux, D. (1989), "An application of the markov chain approach to forecasting cotton yields from surveys," *Agricultural Systems*, 29(4), 357-370.

[71] Mantis, J.H., Saito, T., Grant, W.E., Iwig, W.C., Ritchie, J.T. (1985), "A markov chain approach to crop yield forecasting," *Agricultural Systems*, 18(3), 171-187.

[72] Marra, M.C., Carlson, G.A. (1990), "The decision to double crop: An application of expected utility theory using Stein's theorem," *American Journal of Agricultural Economics*, 72(2), 337-345.

[73] McNew, K., Gardner, B. (1999), "Income taxes and price variability in storable commodity markets," *American Journal of Agricultural Economics*, 81(3), 544-552.

[74] Minnesota Association of Wheat Growers, North Dakota Grain Growers Association and South Dakota Wheat, Inc. (2005), "Will expected grain price returns justify storage costs?," *Prairie Grain Magazine*, 71, http://www.small-grains.org/springwh/MGuide05/storage/storage.html, last accessed on May 17, 2007.

[75] Mitchell, R.A.C., Mitchell, V.J., Driscoll, S.P., Franklin, J., Lawlor, D.W. (1993), "Effects of increased CO_2 concentration and temperature on growth and yield of winter wheat at two levels of nitrogen application," *Plant, Cell and Environment*, 16, 521-529.

[76] Mkhabela, M.S., Mkhabela, M.S., Mashinini, N.N. (2005), "Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR," *Agricultural and Forest Meteorology*, 129, 1-9.

[77] Montgomery, D.C. (1997), *Design and Analysis of Experiments*, Wiley, New York.

[78] National Agricultural Statistics Service, United State Department of Agriculture (2005), "Acreage report," http://usda.mannlib.cornell.edu/usda/nass /Acre//2000s/2005/Acre-06-30-2005.pdf, last accessed on May 17, 2007.

[79] National Agricultural Statistics Service, United State Department of Agriculture (2005), "Agricultural statistics 2005," http://www.usda.gov/nass/pubs /agr05/acro05.htm, last accessed on May 17, 2007.

[80] National Agricultural Statistics Service, United State Department of Agriculture, "Crop production," http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1046, last accessed on May 21, 2007.

[81] National Agricultural Statistics Service, United State Department of Agriculture (1997), "Usual planting and harvesting dates for U.S. field crops," http://www.usda.gov/nass/pubs/uph97.htm, last accessed on May 17, 2007.

[82] National Agricultural Statistics Service, United State Department of Agriculture (2005), "2002 Census of agriculture," http://www.nass.usda.gov/Census_of_Agriculture/, last accessed on May 17, 2007.

[83] National Climatic Data Center, United State Department, http://www.ncdc.noaa.gov/oa/ncdc.html, last accessed on May 17, 2007.

[84] Netz, J.S. (1995), "The effect of futures markets and corners on storage and spot price variability," *American Journal of Agricultural Economics*, 77(1), 182-193.

[85] Oberle, S.L., Keeney, D.R. (1990), "Soil type, precipitation, and fertilizer N effects on corn yield," *Journal of Production Agriculture*, 3(4), 552-527.

[86] Orazem, P., Miranowski, J. (1986), "An indirect test for the specification of expectation regimes," *The Review of Economics and Statistics*, 68(4), 603-609.

[87] Park, S.J., Hwang, C.S., Vlek, P.L.G. (2005), "Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions," *Agricultural Systems*, 85(1), 59-81.

[88] Peng, S., Huang, J., Sheehy, J.E., Laza, R.C., Visperas, R.M., Zhong, X., Centeno, G.S., Khush, G.S., Cassman, K.G. (2004), "Rice yields decline with higher night temperature from global warming," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 9971-9975.

[89] Popovici, V., Thiran, J. (2004), "Pattern recognition using higher-order local autocorrelation coefficients," *Pattern Recognition Letters*, 25, 1107-1113.

[90] Porter, J.R., Gawith, M. (1999), "Temperatures and the growth and development of wheat: A review," *European Journal of Agronomy*, 10(1), 23-36.

[91] Potgieter, A.B., Hammer, G.L., Doherty, A., de Voil, P. (2005), "A simple regional-scale model for forecasting sorghum yield across North-Eastern Australia," *Agricultural and Forest Meteorology*, 132, 143-153.

[92] Ramsay, J.O., Silverman, B.W. (2002), *Applied Functional Data Analysis*, Springer, New York.

[93] Ramsay, J.O., Silverman, B.W. (2005), *Functional Data Analysis*, Springer, New York.

[94] Raychaudhuri, S., Stuart, J.M., Altman, R.B. (2000), "Principal components analysis to summarize microarray experiments: Application to sporulation time series," *Pacific Symposium on Biocomputing*, 5, 452-463.

[95] Roe, T., Antonovitz, F. (1985), "A producer's willingness to pay for information under price uncertainty: Theory and application," *Southern Economic Journal*, 52(2), 382-391.

[96] Ruppert, D., Wand, M.P., Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge, New York.

[97] Sarker, R.A., Talukdar, S., Anwarul Haque, A.F.M. (1997), "Determination of optimum crop mix for crop cultivation in Bangladesh," *Applied Mathematical Modelling*, 21(10), 621-632.

[98] Schölkopf, B., Mika, S., Burges, C.J.C., Knirsch, P., Müller, K.-R., Rätsch, G., Smola, A.J. (1999), "Input space versus feature space in kernel-based methods," *IEEE Transactions on Neural Networks*, 10(5), 1000-1017.

[99] Seif, E., Pederson, D.G. (1978), "Effect of rainfall on the grain yield of spring wheat, with an application to the analysis of adaptation," *Australian Journal of Agricultural Research*, 29, 1107-1115.

[100] Sexauer, B. (1977), "The storage of potatoes and the Maine potatoes futures market," *American Journal of Agricultural Economics*, 59(1), 220-224.

[101] Sheehy, J.E., Mitchell, P.L., Ferrer, A.B. (2006), "Decline in rice grain yields with temperature: Models and correlations can give different estimates," *Field Crops Research*, 98, 151-156.

[102] Shibayama, M. (1991), "Estimating grain yield of maturing rice canopies using high spectral resolution reflectance measurements," *Remote Sensing of Environment*, 36(1), 45-53.

[103] Shonkwiler, J.S. (1982), "An empirical comparison of agricultural supply response mechanisms," *Applied Economics*, 14(2), 183-194.

[104] Shonkwiler, J.S., Emerson, R.D. (1982), "Imports and the supply of winter tomatoes: An application of rational expectations," *American Journal of Agricultural Economics*, 64(4), 634-641.

[105] Silveira de Jasa, M.I. (1986), A markov chain model for cotton yield forecasting, Ph.D. Dissertation, Texas A&M University, Texas.

[106] Stephens, D.J., Walker, G.K., Lyons, T.J. (1994), "Forecasting Australia wheat yields with a weighted rainfall index," *Agricultural and Forest Meteorology*, 71(3-4), 247-263.

[107] Sugar, C.A., James, G.M. (2003), "Finding the number of clusters in a data set: An information theoretic approach," *Journal of the American Statistical Association*, 98, 750-763.

[108] Swaney, D.P., Jones, J.W., Mishoe, J.W., Baker, F. (1986), "A combined simulation-optimization approach for predicting crop yields," *Agricultural Systems*, 20(2), 133-157.

[109] Tibshirani, R., Walther, G., Hastie, T. (2001), "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society, B*, 63, 411-423.

[110] Tomek, W.G. (2000), "Commodity prices revisited," Staff Paper, Cornell University, New York.

[111] Tomek, W.G., Gray, R.W. (1970), "Temporal relationships among prices on commodity futures markets: Their allocative and stabilizing roles," *American Journal of Agricultural Economics*, 52, 372-380.

[112] Topaloglou, N. (2004), A stochastic programming framework for international portfolio management, Ph.D. Dissertation, University of Cyprus, Cyprus.

[113] Tronstad, R., Taylor, C.R. (1991), "Dynamically optimal after-tax grain storage, cash grain sale, and hedging strategies," *American Journal of Agricultural Economics*, 73(1), 75-88.

[114] University of Illinois (2005), "Grain farm returns and costs," http://www.farmdoc.uiuc.edu/manage/grain_farm_returns_costs.pdf, last accessed on May 17, 2007.

[115] University of Illinois (2001), "The leasing forum," http://www.farmdoc.uiuc.edu/farmland/publications/Leasing_Forum/leasing_forum.html, last accessed on May 17, 2007.

[116] University of Missouri-Columbia (1997), "No-tillage and conservation tillage: Economic considerations," *MU Guide*, G355, http://extension.missouri.edu/explorepdf/agguides/agecon/g00355.pdf, last accessed on May 17, 2007.

[117] van Elderen, E. (1980), "Models and techniques for scheduling farm operations: A comparison," *Agricultural Systems*, 5(1), 1-17.

[118] Vogel, F.A., Bange, G.A. (1999),"Understanding USDA crop forecasts," United State Department of Agriculture, Miscellaneous Publication No.1554, http://www.nass.usda.gov/Data_and_Statistics/pub1554.pdf, last accessed on May 17, 2007.

[119] Walker, G.K. (1989), "Model for operational forecasting of Western Canada wheat yield," *Agricultural and Forest Meteorology*, 44(3-4), 339-351.

[120] Wheeler, T.R., Craufurd, P.Q., Ellis, R.H., Porter, J.R., Prasad, P.V.V. (2000), "Temperature variability and the yield of annual crops," *Agriculture, Ecosystems & Environment*, 82, 159-167.

[121] Wilcox, A., Perry, N.H., Boatman, N.D., Chaney, K. (2000), "Factors affecting the yield of winter cereals in crop margins," *Journal of Agricultural Science*, 135(4), 335-346.

[122] Wilks, D.S., Pitt, R.E., Fick, G.W. (1993), "Modeling optimal alfalfa harvest scheduling using short-range weather forecasts," *Agricultural Systems*, 42(3), 277-305.

[123] Working, H. (1942), "Quotations on commodity futures as price forecasts," *Econometrica*, 10, 39-52.

[124] World Agricultural Outlook Board, United State Department of Agriculture, "World agricultural supply and demand estimates," http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1194, last accessed on May 11, 2007.

[125] Yao, F., Müller, H.G., Wang, J.L. (2005), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577-590.

[126] Yu, L., Ji, X., Wang, S. (2003), "Stochastic programming models in financial optimization: A survey," *Advanced Modeling and Optimization*, 5(1), 2003.

[127] Zhang, P., Anderson, B.T., Myneni, R. (2006), "Monitoring 2005 corn belt yields from space," *EOS, Transactions American Geophysical Union*, 87, 150.

VITA

Nantachai Kantanantha was born in 1976 in Bangkok, Thailand. He received his Bachelor's Degree in Industrial Engineering from Chulalongkorn University, Thailand, in 1997. From 1997 to 2000, he worked at Research and Process Development Department, Kasikorn Bank, the third largest bank in Thailand. He worked as an internal consultant for re-engineering the old processes of other departments. In the end of 1999, he received a scholarship from the Royal Thai Government to pursue his study in the U.S. for Master's and Ph.D. degrees in Industrial Engineering. In August 2000, he decided to join the School of Industrial and Systems Engineering at Georgia Institute of Technology and received his Master of Science in Industrial Engineering in December 2001. He then joined the Ph.D. program at the same institute with a concentration on economic decision analysis. He received his Ph.D. in August 2007.